Differential splicing in lymphoma

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr. rer. nat.
im Fach Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Humboldt-Universität zu Berlin

von

Dipl.-Inf. Karin Zimmermann

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Ulf Leser
2. Prof. Michael Hummel
3. Prof. Kay Nieselt

eingereicht am: 26.06.2017
Tag der mündlichen Prüfung: 18.12.2017

Zusammenfassung

Alternatives Spleißen ist ein wesentlicher Mechanismus, um Proteindiversität in Eukaryoten zu gewährleisten. Gewebespezifität sowie entwicklungsrelevante Prozesse werden unter anderem massgeblich davon beeinflusst. Aberrante (alternative) Spleißvorgänge können wiederum zu veränderten Proteinisoformen führen, die verschiedenste Krankheiten wie Krebs verursachen oder zu veränderter Medikamentenwirksamkeit beitragen können. In dieser Arbeit untersuchen wir differentielles Spleißen im Kontext von Krebserkrankungen. Dazu betrachten wir drei Aspekte, die uns wichtig erscheinen.

Der erste Teil dieser Arbeit beschäftigt sich mit dem systematischen Vergleich verschiedener Methoden für die Detektion von differentiellem Spleißen in Exon-Array-Daten. Mehrere Methoden wurden für diese Aufgabe entwickelt, die Übereinstimmung der jeweiligen Ergebnisse ist jedoch oft gering. Darüber hinaus gibt es bis heute keine Evaluierung dieser Methoden bezüglich des Einflusses bestimmter Datenparameter auf die Ergebnisgüte. Aus diesem Grund haben wir artifizielle Daten generiert, um Parameter wie Expressionsintensität oder die Anzahl der differentiell gespleißeten Proben pro Gruppe zu modellieren. Zusätzlich haben wir alle Methoden auf experimentell validierte Datensätze angewandt. Mit Hilfe unseres Ansatzes identifizieren wir Methoden, die über alle Szenarien hinweg robuste Ergebnisse liefern, und ermitteln bestimmte Datenparameter, die die Ergebnisgüte sowie die Qualität der angewandten Methoden beeinflussen.

Im zweiten Teil identifizieren wir Spleiß-regulatorischer Proteine, die für die beobachteten Spleissveränderungen zwischen Krebs, hier Lymphomen, und einer Kontrolle, verantwortlich sein könnten. Erkenntnisse derartiger Zusammenhänge ermöglichen ein besseres Verständnis der zugrundeliegenden Pathomechanismen und folglich bessere Therapiemöglichkeiten. Zu diesem Zweck stellen wir eine von uns entwickelte Methode basierend auf einem Netzwerkansatz vor. Hierbei werden Spleißfaktoren und differentiell gesplicete Exons in ein Netzwerk integriert und anschliessend anhand der Unterschiede in ihrer Zentralität geordnet. Wir untermauern unseren Ansatz, indem wir die Platzierung differentiell exprimierter Regulatoren unter den Topkandidaten untersuchen, und diskutieren anhand ausgewählter Beispiele die potentiellen Einflussursachen für nicht differentiell exprimierte Kandidaten.

Im dritten Teil analysieren wir die Vergleichbarkeit zweier Datentypen, generiert durch unterschiedliche Technologien, in Bezug auf die Detektion von differentiellem Spleißen. Eine Abschätzung dieser Vergleichbarkeit ist von hohem Interesse, da RNA Sequenzierung Exon Arrays nach und nach ablöst. Vergleiche dieser Art sind selten und aufgrund geringer Fallzahlen oder der Verwendung seltener Microarrays wenig aussagekräftig. Um diese Aufgabe zu lösen, entwickeln wir einen Multiebenenansatz, der den Vergleich nicht auf die Ebene des differentiellen Spleißens reduziert, sondern vorhergehende Ebenen miteinbezieht. Um eine Beeinträchtigung der Vergleichbarkeit durch die Verwendung unterschiedlicher Methoden für die Detektion von differentiellem Spleißen zu vermeiden, wenden wir Methoden an, die für beide Technologien geeignet sind. Die Anwendung unseres Ansatzes auf zwei Datensätze identifiziert ähnliche Trends in der Vergleichbarkeit bei einer sich unterscheidenden Gesamtkonkordanz.

Abstract

Alternative splicing is a crucial mechanism in eukaryotes, which provides an ample protein diversity that is necessary for maintaining an organism by, for instance, establishing tissue diversity and contributing to developmental processes. In contrast, aberrant (alternative) splicing may lead to altered protein isoforms contributing to cellular malfunctioning, diseases such as cancer as well as altered susceptibility towards drug treatment. In this thesis, we study differential splicing in cancer, i.e. splicing changes observed between cancerous and control tissues. More precisely, we seek to identify methods best suited for the detection of differential splicing, we investigate regulatory factors potentially causal for the splicing changes observed, and we study the comparability of two data types obtained from different technologies with respect to differential splicing detection.

The first part of the thesis assesses the performance of methods for detecting differential splicing from exon arrays. Several methods exist for this task, but their results are often of low concordance. Moreover, no comprehensive evaluation has been conducted so far to examine global data parameters and their potential influence on results and method performance. To this end, we generated artificial data to model different variables, such as expression intensity or the percentage of differentially spliced samples per group. Additionally, we applied all methods compared to validated experimental data. Overall, our evaluation indicates methods that perform robustly well across artificial and experimental data and identifies parameters impacting result performance.

The second part aims at identifying regulatory factors responsible for splicing changes observed between cancer, namely lymphoma, and healthy tissue. Determining the underlying causes for these events might elucidate the pathological mechanism of the disease and thus, provide options for targeting carcinogenic enhancers. To this end, we develop a novel, network based approach which first integrates differentially spliced exons with splicing regulatory proteins (splicing factors), using transcriptomics data, and then ranks splicing factors according to their potential involvement in cancer. We strengthen our approach by assessing the enrichment of differentially expressed splicing factors amongst our predicted causal regulators and discuss selected examples of alterations potentially influencing splicing factors highly ranked in our approach, which are not differentially expressed.

Third, we compare differential splicing detection based on RNA sequencing and exon array data. As RNA sequencing is gradually replacing microarrays, the comparability with respect to this task is of crucial interest. While gene based comparisons show a high result concordance, studies on differential splicing are more seldom and lack explanatory power due to small sample sizes or rarely used microarrays. We address this task by developing a multi-level comparison framework, elucidating comparability on several levels including, for instance, probe and gene level. Additionally, we aim at improving differential splicing comparability by considering two differential splicing detection methods applicable to both, RNA sequencing and exon array data, to avoid method inherent bias. We apply our multi-level framework to two data sets, leading, despite varying overall concordance, to similar trends in comparability.

Acknowledgments

This thesis would not have been possible without the encouragement and support of many people.

First of all, I thank my supervisor, Prof. Ulf Leser, I am grateful for his constant guidance and support through quite some years. His encouragement and his scientific advice have been indispensable for this work. Also, I am very thankful for the opportunity to work in his group, which provided an exceptionally friendly and fruitful atmosphere.

I am thankful to Prof. Michael Hummel, who supported me with guidance on the biological aspects of my work, and provided me with the opportunity to work on various interesting and exciting research topics. I also thank Apl. Prof. Kay Nieselt, who paved the way for this work by awaking my interest for this field during my studies.

The last years wouldn't have been that 'awesome' without all the wonderful people at WBI. Thanks to all members throughout the years for an interesting and open-minded atmosphere, for eye-opening discussions on and off scientific topics (I now know how to disembowel a bear in under 15 minutes) and amusing social activities.

I also would like to thank the people at Charité for a wonderful time. They provided me with a different perspective and strengthened the enthusiasm for my work trough fruitful cooperations. I also want to thank them for helpful advice regarding my thesis.

Special gratitude is dedicated to Samira, Martin and Phil for proof-reading large parts of my thesis. I am also especially thankful to Sven, who not only tolerated some ups and downs regarding my thesis, but also helped me grow through them. Finally, I thank my parents and sisters for their unconditional love and support throughout my life.

This work was funded by the DFG through TRR54 as well as through the EU project BiobankCloud. Finalization of my work has been made possible through the grant provided by the Caroline von Humboldt Stiftung. I am highly grateful for this support.

Contents

1 Introduction

The finalization of the Human Genome Project [Collins et al., 2003] in 2003 has revolutionized genomics, and, in consequence, biology and medicine [Collins, 1999, Collins et al., 2003]. Knowledge about the nucleotide sequence constituting a human genome enabled a fundamentally new approach to the investigation of all cellular processes giving rise to several high-throughput technologies. As of that moment, the study of biological entities, such as genes, emerged from single to all-in-parallel [Kononen et al., 1998, Mardis, 2008]. Such global approaches allowed for a new understanding about individual entities, and, more importantly, their interdependence as well as their interplay and regulation [Mardis, 2008, Sorek and Cossart, 2010].

Different layers are involved in maintaining the subtle equilibrium of life. The genome encodes this information, which is first transcribed to mRNA and then translated to the active cellular components, the proteins. A major control instance for this essential process is the epigenome, defining chemical changes of the DNA and histone proteins providing or restricting access to the transcription machinery. These different layers interact and regulate each other [Luco et al., 2011, Fu et al., 2013]. Thus, aberrant behaviour in any of these layers may lead to malfunction of an organism resulting in disease or unusual therapeutic responses.

Various large-scale 'omics' technologies have been developed as investigatory tools to analyze the individual biological layers and regulatory processes at large. Transcriptomics captures the amount of mRNA present for all transcribed entities, such as genes, in a sample at a certain time point. It is widely applied to determine differentially expressed genes between two or more conditions. This approach has led over decades to valuable insights changing the understanding of biological mechanisms as well as therapeutic approaches successfully. Gene signatures predict susceptibility to therapy [Van't Veer et al., 2002], survival [Van De Vijver et al., 2002], and enable classification and class discovery of disease subtypes [Calon et al., 2015, Golub et al., 1999, Tibshirani et al., 2002].

However, the early one-gene-one-protein hypothesis has been overcome decades ago, as the observed number of genes in eukaryotes, about 25,000 genes in human, is not able to provide the physiological complexity necessary to maintain an organism [Graveley, 2001]. The major mechanism responsible for this augmentation in protein diversity is alternative splicing. Splicing is an inevitable step in the transcription of multi-exon genes, responsible for the excision of introns. The remaining parts, exons, are the information-coding genomic sequences that are joint (all or some) into the final transcript, which is then subjected to translation. Accordingly, splicing is also responsible for variation of the sequences included into the final transcript. This process, referred to as alternative splicing, leads to an increase in complexity of the functional proteome by varying

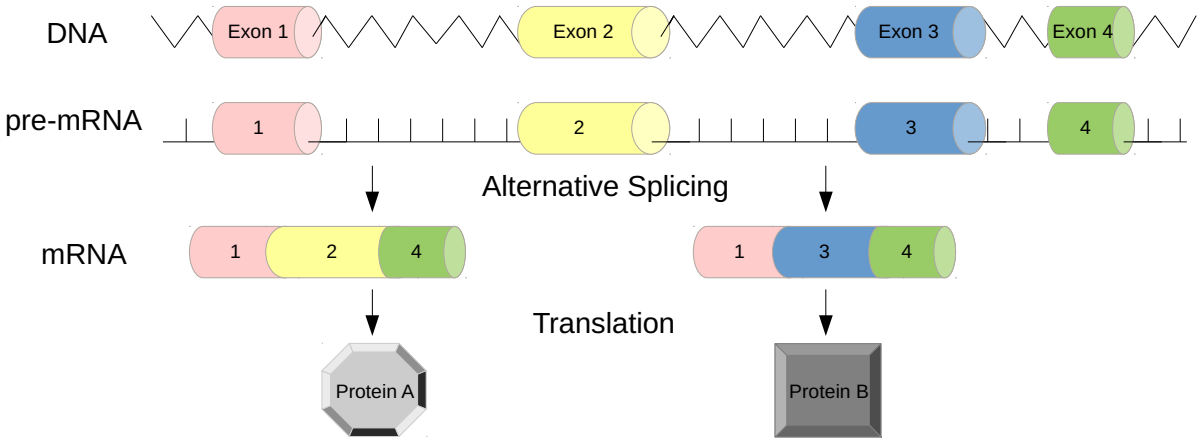


Figure 1.1: **Alternative splicing.** The information constituting a gene is transcribed from the DNA to a pre-mRNA. Alternative Splicing leads to different mRNAs and subsequently to different protein isoforms.

Erroneous splicing can lead to protein products with altered function. While this can be an evolutionary advantage in few cases, most of the outcomes are physiologically disadvantageous and lead to severe diseases such as neuro-degenerative disorders [Golde et al., 1990, Meshorer and Soreq, 2006] and cancer [Piekielko-Witkowska et al., 2010, Watson et al., 2013, Lenzken et al., 2013, Fackenthal and Godley, 2008]. A biologically well-documented example of altered protein function due to a different isoform of a gene is discovered for BCL-X. This gene is a family-member of apoptotic regulators. A short, BCL-X_S, and a long isoform, BCL-X_L, is known. While BCL-X_L has an anti-apoptotic effect, BCL-X_S displays the adverse function by promoting apoptosis. Accordingly, the two isoforms of this gene are of high interest as therapeutical targets [Akgul et al., 2004].

Regulation of splicing is mainly accomplished by two types of elements (see Figure 1.2). The first ones, referred to as cis-regulatory elements, are nucleotide sequences located in the DNA, which interact with trans-regulatory elements, i.e. proteins involved in the process of splicing and its regulation. These elements can in turn be influenced by general superordinate layers.

Several mechanisms can contribute to patho-physiological splicing. For instance, about 30% of all disease causing mutations unfold their malicious impact by altering the splicing pattern [Xi et al., 2008]. Furthermore, growing evidence in the last years suggested a central role of epigenetics in the regulation of alternative splicing. This also pinpointed to aberrant splicing events linked to epigenetic alterations in cancer [Li et al., 2015, Zhao et al., 2017, Salton et al., 2014].

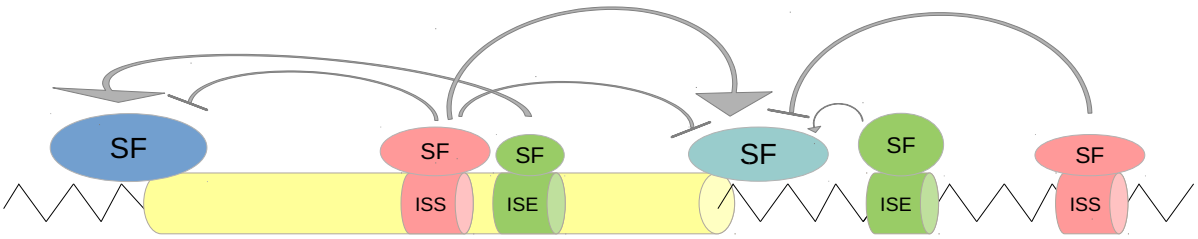


Figure 1.2: **Regulation of alternative splicing** is controlled by cis-acting splicing regulatory elements (SRE) and trans-acting splicing factors. Cis-splicing regulatory elements can be distinguished into intron or exon splicing silencers or enhancers (ISE, ISS, ESE, and ESS) depending on their relative location and their impact on the splicing event. Splicing factors are recruited by corresponding cis-regulatory factors and promote or inhibit splice site recognition.

Recent studies emphasized the crucial role of aberrant splicing in cancer [Oltean and Bates, 2014, Sveen et al., 2015, Dvinge et al., 2016, He et al., 2014]. Yet, the detailed instances of aberrant (differential) splicing specific for different cancer types or even general patterns of aberrant splicing in cancer together with the regulatory mechanisms leading to such aberrant splicing still have to be elucidated.

The detection of different splicing patterns between conditions on the transcriptomic level can be assessed with two major technologies, (exon) microarrays and RNA sequencing. The first generation of high-throughput transcriptome technologies were microarrays [Schena et al., 1995]. They capture the amount of mRNA of interest by hybridization of complementary probes attached to a small chip. This technique is still widely used, as microarrays have proven themselves worthwhile [Van't Veer et al., 2002, Tibshirani et al., 2002, Calon et al., 2015]. However, they have several drawbacks, the most obvious being the fact that only known entities represented on the microarray can be interrogated. With advances in technology, next-generation high-throughput approaches emerged. Here, identification and quantification of the interrogated genomic entities is based on sequencing.

While next-generation sequencing (NGS) is applicable to the same scenarios as microarrays, it additionally enlarges and complements their repertoire mainly through an unbiased view as no prior knowledge of the sequence is needed. Additionally, NGS provides a more fine-grained resolution on sequence copies, as it is designed to be capable of reporting every single transcript encountered as opposed to microarrays, where expression is represented in an analog rather than discrete way.

1.1 Aim

Accurate prediction of the differential splicing (DS) events distinguishing two conditions such as cancer and control tissue together with the regulatory malfunctioning causing

such aberrant splicing is essential for understanding the disease pathology and identifying therapeutical targets.

A plethora of methods for the detction of differential splicing from exon arrays exists. Nevertheless, the overlap in their prediction of DS events is at times low [Bisognin et al., 2014]. Additionally, no comprehensive assessment to susceptibility of global data parameters potentially influencing result performance has been conducted so far. Yet, these influences can be vital for the adequate selection of a method for a defined study aim such as biomarker detection [Hartwell et al., 2006]. A central aim of our work is thus to (1) study and compare algorithms for the identification of aberrant splicing events characteristic for a certain condition together with their performance according to different measures and their suceptibility to several global data parameters. Identifying the method most adequate for the prediction of differential splicing events according to performance based on the given data parameters and the defined study aim may lead to the most reliable and appropriate predictions.

Several layers might contribute to the malfunctioning of splicing regulators. While in principle different omics technologies might be used to interrogate potential causes for the aberrant regulatory impact on splicing, such an approach is associated with a tremendous time and cost investment. Additionally, no gold standard for the integrative analysis of different omics data sets has been established so far. We thus (2) aim at developing a novel algorithm for the identification of the regulatory elements causal for the splicing changes observed without restricting potential alterations in regulatory elements to the transcriptome level, yet, while only using transcriptomic data.

Given the advances NGS technologies provide, they are gradually replacing microarrays. Thus, comparability of the two technologies is of great interest to determine whether research results can be reproduced. For the gene expression level, this comparison has been conducted extensively, however, an evaluation based on differential splicing is rather seldom and lacks explanatory power due to small sample size or rarely used microarrays [Raghavachari et al., 2012, Bradford et al., 2010]. Therefor, we (3), aim at identifying comparability of RNA sequencing data to exon array data with respect to differential splicing detection.

1.2 Contributions

Our aim in this work is to develop methods to improve the understanding of splicing in cancer, especially in lymphoma. Only a detailed understanding of occurrences of differential splicing, i.e. differences in splicing between conditions, as well as the circumstances that provoke this occurrence can lead to a holistic understanding and thus a targeted therapeutic approach based on specific isoform expressions or causal regulators. To this end, we address three questions we believe to be most relevant for differential splicing and its role in cancer.

- First, we tackle the problem of detecting differential splicing based on exon arrays. While a variety of methods for this task exist, result sets differ substantially

and no evaluation based on different global data features influencing results has been presented to date. We perform a comprehensive study to identify the best method(s) for a given scenario based on artificial and real data sets comprising several aspects potentially influencing detection of differential splicing events. We evaluate and discuss results with regard to the mathematical background of the methods and give recommendations on which method to use for a certain scenario.

- Second, we aim at elucidating regulatory changes responsible for the differential splicing observed. To this end, we integrate differential splicing events observed with potential regulators, i.e., splicing factors (SF), in an expression correlation network. Hereby, we obtain a list of candidate splicing factors. Our approach is designed to reflect not only transcription induced changes, but also changes observed in other layers such as epigenetics or genomics even though using only one technology, i.e. transcriptomics. We corroborate our result by showing a high ranking of differentially expressed splicing factors in our candidate lists. Furthermore, we investigate and discuss potential sources of influence on the identified splicing factors.

- Third, we propose a framework for the elucidation of the concordance between RNA sequencing and exon arrays with respect to differential splicing prediction. While several works show comparability of results on gene level, result comparison on differential splicing basis so-far are rare and lack explanatory power. To this end, we implement a multi-level framework allowing for comparison of the two technologies on various levels. We also study result comparability by using the same methods for differential splicing detection on both data types. Our approach shows substantial comparability throughout data sets. Additionally, we determine the influence of different factors on result concordance.

1.3 Outline of Thesis

This section gives a brief overview on structure and content of this work. We highlight the different aims, the solutions developed to achieve our tasks as well as a short overview on the results in a chapter-based manner.

Chapter 2 gives an introduction to the two main aspects relevant throughout this work: First, the mechanism of (alternative) splicing, its regulation as well as its implications for diseases such as cancer. Second, we introduce high-throughput technologies for the quantification of transcription in the context of differential splicing. We give detailed insights on the characteristics of exon array and RNA sequencing derived data and discuss their strengths and weaknesses. Last, we describe the key data set of this thesis used throughout most of this work.

Chapter 3 introduces several methods for differential splicing detection based on exon array data and motivates the need for detailed, thorough comparisons based on controlled data scenarios. Data simulation as well as experimental influences are introduced. Results on artificial data as well as for two q-PCR validated data sets are presented and

discussed with respect to the mathematical method background as well as to the scenarios created. Finally, we give recommendations for the choice of a differential splicing detection method for different application scenarios.

Chapter 4 describes a novel, network based method for the detection of regulatory elements causal for the splicing changes observed in a condition. To this end, we perform an integrated analysis on altered entities, i.e. differentially spliced exons together with potential regulators, i.e. splicing factors using a network based approach. We evaluate our method by investigating the ranking of differentially expressed splicing factors amongst our network-based top candidates for aberrant regulators. We contextualize our findings with literature research and published results in potentially causal layers.

Chapter 5 is dedicated to the comparison of RNA sequencing and exon array derived differential splicing detection. As microarrays are gradually replaced by RNA sequencing, we thoroughly evaluate concordance of the differential splicing events detected using a multi-level approach designated to elucidate changes on several levels. We tackle potentially induced method bias by the application of the same methods on both data types and evaluate the impact of different factors on result concordance.

Chapter 6 summarizes this thesis, highlights the central contributions of this work and addresses possible future directions.

1.4 Own Prior Work and Contribution

Chapter 3 describes the comparative assessment of differential splicing detection methods originally presented in [Zimmermann et al., 2015] and based on previous work from [Zimmermann and Leser, 2010]. The contributions to this work can be attributed to the authors as follows: For [Zimmermann and Leser, 2010], Karin Zimmermann and Ulf Leser conceived the research, Karin Zimmermann wrote and Ulf Leser revised the manuscript. In [Zimmermann et al., 2015], Karin Zimmermann and Ulf Leser conceived the research. Karin Zimmermann carried out the experiments, analyzed the results and drafted the manuscript. Marcel Jentsch developed KLAS, a method for detection of differential splicing, Marcel Jentsch and Axel Rasche implemented the differential splicing detection methods. Axel Rasche, Michael Hummel, and Ulf Leser helped to revise the manuscript and to interpret the data.

2 Biological and Technical Background

In the last decades high-throughput technologies have revolutionized the way complex genomic questions can be answered [Li et al., 2014]. By targeting all elements of the genome at a time and in parallel, a way more fast, cheap and objective investigation is ensured. On the transcriptome level, this allows for the quantification of exon expression using technologies such as exon arrays and high-throughput sequencing. This more fine-grained resolution, as opposed to the quantification on gene level, enables the detection of alternative splicing. Discovered in the early 70s, the mechanism of alternative splicing provides the major source of transcript and therefor protein diversity which is necessary to maintain eukaryotic physiology [Nilsen and Graveley, 2010]. As however the mechanism is highly complex and depends on various players and their exact abundance, numerous diseases are known to be associated with aberrant splicing [Golde et al., 1990, Meshorer and Soreq, 2006, Piekielko-Witkowska et al., 2010, Watson et al., 2013, Lenzken et al., 2013]. Thus, differences in isoforms between conditions, i.e. variations of a protein due to different splicing, are of high therapeutic interest.

In this chapter we will describe the mechanism of alternative splicing along with its regulation and association with diseases. Furthermore, we introduce high-throughput technologies such as microarrays and high-throughput sequencing that are crucial for detecting alternative splicing.

2.1 Alternative Splicing

During protein biosynthesis a gene is translated into a protein (Figure 2.1). First, the genetic code is transcribed into a pre-mRNA which is an exact copy of the genetic sequence. This sequence contains the exons - the information coding part - as well as intronic regions, usually not part of the final template. To obtain the mature mRNA, all sequences which are not intended to be part of the final template are removed, i.e. spliced out. The resulting mRNA is now translated into the corresponding amino acid sequence by the ribosomal machinery of the cell which in turn is folded into the functional protein. However, the process of splicing is not static, i.e. a gene can be translated into different mRNAs and thus into different proteins by excluding exons from the final transcript. This process is called alternative splicing and provides the variety of proteins necessary to maintain the proper functioning of an eukaryotic organism.

Since advances in large-scale technologies such as microarrays and high-throughput sequencing are substantially alleviating genomic research, the extent of the prevalence and thus the impact of splicing becomes more and more clear. According to recent studies over 95% of all multi-exon genes are alternatively spliced [Wang et al., 2008]. This

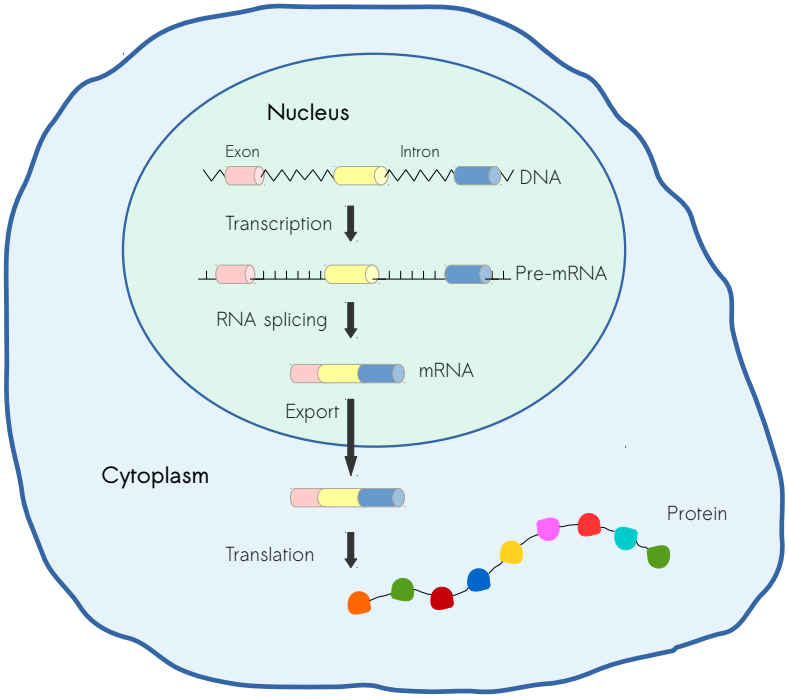


Figure 2.1: **Protein biosynthesis** includes transcription of entities encoded in the DNA and their translation to functional proteins.

underlines the fundamental importance of alternative splicing for the establishment of the protein diversity necessary for eukaryotes. Being such an integral part of gene regulation, the mechanism of alternative splicing is a highly relevant subject of investigation in physiological scenarios as well as in aberrant conditions such as diseases.

Here, we will discuss the importance of alternative splicing for eukaryotes, its mechanism, types, and regulation. We will also give an overview on disease associated splicing with a special focus on lymphoma. We will use the term *alternative splicing* with the meaning described above, i.e. referring to the general mechanism. In contrast, *aberrant splicing* refers to non-physiological instances of splicing as, for instance, observed in diseases. By *differential splicing* we refer to splicing events which differ between two groups being analyzed. While these terms might overlap in their meaning, we adhere to a context-dependent usage.

2.1.1 The Importance of Alternative Splicing for Eukaryotes

When the human genome was decoded, the number of genes encountered was somewhat surprising. Roughly 25,000 genes constitute the genome. For an organism as complex as the human a much higher number of proteins is necessary to maintain the intricate processes underneath. The increase in protein diversity is achieved by alternative splicing. This mechanism allows for a number of protein isoforms higher by one magnitude [Hu et al., 2015].

On an evolutionary scale, alternative splicing can provide faster adaption to new environmental challenges. In a prokaryotic scenario the usual point mutation is a rather slow development towards a new function. For eukaryotes, the in- or exclusion of exons by, for instance, a point mutation, provides a faster new feature as exons are already 'approved' components.

Alternative splicing is known to play a central role in several physiological scenarios. It affects tissue diversity [Grosso et al., 2008, Calarco et al., 2011, Yeo et al., 2004], development [Giudice et al., 2014, Revil et al., 2010, Barberan-Soler and Zahler, 2008], protein stability [Jensen and Whitehead, 2001, Sakurai et al., 2001], post-translational modifications [Naro and Sette, 2013, Liu et al., 2013, Fackenthal and Godley, 2008], enzymatic activity [Li and Koromilas, 2001], gene expression regulation [Heinzen et al., 2008], apoptosis [Schwerk and Schulze-Osthoff, 2005], DNA damage repair [Lenzken et al., 2013] intracellular localization [Indraccolo et al., 2002, Koch et al., 2001], and binding properties [Birikh et al., 2003]. Correspondingly, the mechanism is indispensable for all aspects of eukaryotic maintenance. Sometimes, alternative splicing can even lead to a binary function switch, as in the case of a pro- and anti-apoptotic isoform of a gene [Akgul et al., 2004, Thorsen et al., 2008].

2.1.2 The Mechanism of Splicing

The basic mechanism of splicing is accomplished by the so-called spliceosome. This conglomerate of five small nuclear ribonucleic proteins (snRNPs) catalyzes the excision of introns from the pre-mRNA and mediates the union of the exons. The five snRNPs, U1, U2, U4, U5, and U6 , are known as the *major spliceosome*, responsible for canonical splicing, which is active in the nucleus and accomplishes the splicing of a vast majority of the introns. Yet, a *minor spliceosome* exists, accounting for a small number, only about 1%, of introns, which, except for U5, depend on different snRNPs namely U11, U12, U4atac, and U6atac [Patel and Steitz, 2003]. For the assembly of the spliceosome the additional factors U2 small nuclear RNA auxiliary factor 1 (U2AF35), U2AF2 (U2AF65) and SF1 are required [Black, 2003, Matlin et al., 2005].

Three sites in an intron are fundamental for the mechanism of splicing. Two of them, the splice sites, are found at the 5' end and the 3' end of the intron consisting of the nucleotides GU and AG respectively. The third relevant site, the branch point, is located upstream from the 3' end of the intron and contains an adenine, while flanking nucleotides are somewhat variable. Subsequently, the pyrimidine tract - a series of pyrimidines - is found.

The actual splicing process is constituted of several steps with different active complexes [Wahl et al., 2009, Will and Lührmann, 2011, Kornblihtt et al., 2013]. An overview is given in Figure 2.2. First, complex E is formed. To this end, U1 binds to the GU sequence at the 5' splice site of the intron, SF1 attaches to the branch point while U2AF1 binds to the 3' splice site and U2AF2 to the polypyrimidine tract. Subsequently, the pre-spliceosome, known as complex A, is assembled. By hydrolyzation of ATP, U2 binds to the branch point. A trimer consisting of the snRNPs U4, U5 and U6 binds, giving rise to the pre-catalytic spliceosome (complex B), which is then, by the release of U1

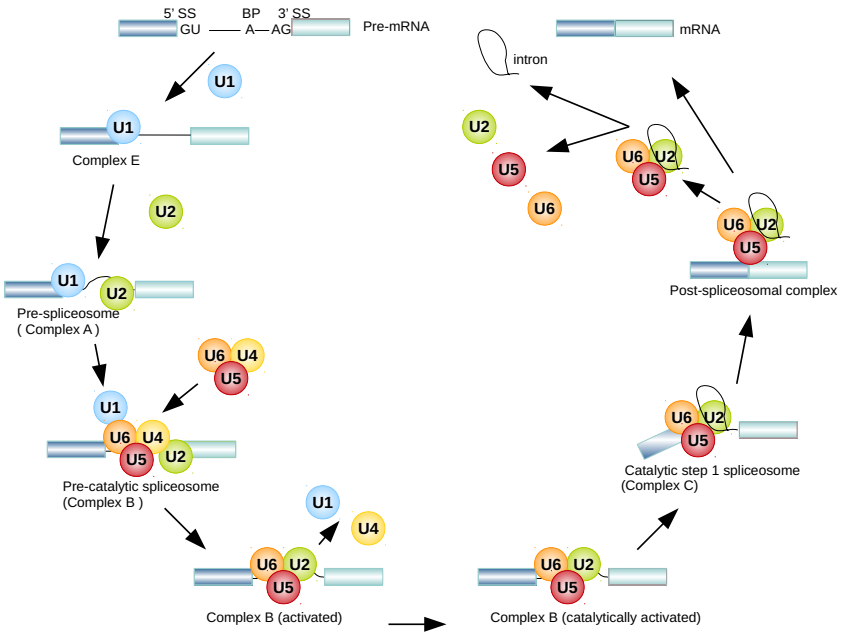


Figure 2.2: **The spliceosome** catalyzes the excision of introns from the transcribed pre-mRNA and mediates the concatenation of exons for the final mRNA.

and U4, turned into a catalytically active complex B*. Hence, the first of two transesterification steps, catalyzed by U2 and U6 can take place. The intron lariat is formed, a branched RNA, which results from the nucleophilic attack by the 2'OH group of the adenosine in the branch point on the 5' splice site (catalytic spliceosome, complex C). The second transesterification step produces the spliced mRNA and the excised intron lariat by attack of the 3'OH group of the upstream exon on the 3' splice site. As soon as the exons are joint to the final mRNA, U2, U5 and U6 are released and the lariat is degraded.

2.1.3 Types of Alternative Splicing

According to current knowledge, not all exons are part of the variety-providing pool. While some exons are contained in all transcripts produced from a gene, known as *constitutive exons*, others may or may not be included in the final transcript and are thus responsible for the observed protein diversity. These exons are called *alternative exons.* Most of the alternative splicing events observed can be classified into the subsequent categories (Figure 2.3) [Keren et al., 2010]:

a *exon skipping* - one or more exons are excluded from the final transcript

b *alternative 3' splice site selection* - the exon is shortened or elongated on the 3' splice site

c *alternative 5' splice site selection* - the exon is shortened or elongated on the 5' splice site

d *intron retention* - the intron is included in the transcript

e *mutually exclusive exons* - only one of two exons is included in the transcript

f *alternative promoters* - the transcript starts at a different position

g *alternative polyA* - the transcript ends at a different position

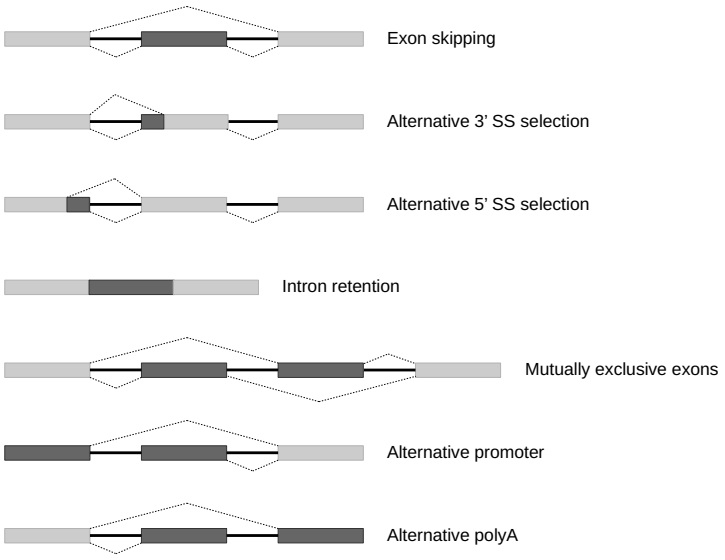


Figure 2.3: **Different types of alternative splicing** exist. Exons can be skipped, at times mutually exclusive, and different 3' or 5' splice sites can be used. Transcript alteration can also occur due to the use of alternative promoters or polyadenylation.

While the use of alternative promoters and alternative adenylation sites are not actually splicing events, they can usually be detected with similar methods using high-throughput sequencing or adequate microarrays.

Some types of events are much more frequent than others [Sammeth et al., 2008]. In human, exon skipping is by far the most frequent event (see Figure 2.4). In most cases, splicing occurs in one transcript, nevertheless trans-splicing is known, where exons from different pre-mRNA transcripts are ligated together [Iwasaki et al., 2009].

Not all of these events can be detected with the use of exon arrays. By design, only de- or inclusion of exons can be detected.

2.1.4 Regulation of Alternative Splicing

Besides the elementary splicing signals, the two splice sites at the 5' and 3' end of an intron as well as the branch point located in close proximity upstream to the 3' splice site, multiple sites and proteins exist, which control splicing in a complex and interdependent manner. These regulatory elements are referred to as cis- and trans-acting factors, depending on whether they are located in the genome (cis) or whether

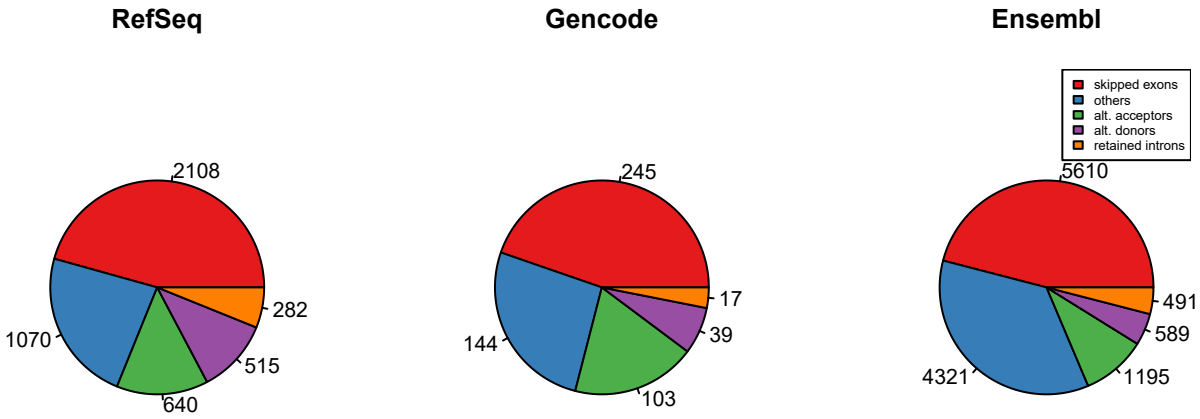


Figure 2.4: **Frequency of human alternative splicing types** according to Gencode, RefSeq and Ensembl [Sammeth et al., 2008].

it is a matter of the corresponding regulatory proteins (trans), respectively [Wang and Burge, 2008]. Trans-acting factors are also referred to as splicing factors, which is the term of detonation used in this work.

Generally, both types of elements can act as promoters or inhibitors of splicing [Barash et al., 2010]. In the case of trans-acting factors, they are denoted as activators or repressors, when referred to cis-acting elements, enhancers and silencers.

The effect of cis-acting elements depends on their sequence, the proximity to the exon/intron of interest as well as the abundance and type of trans-acting factors present. A cis-acting sequence can, for instance, repress or enhance the excision of a nearby intron and therefor act as an intron splicing silencer (ISS) or intron splicing enhancer (ISE). Similarly, the inclusion of exons can be controlled by cis-acting factors operating as exon splicing enhancers (ESE) or exon splicing silencers (ESS).

The same principle holds for trans-acting factors. Depending on the cis-acting factors and additional trans-acting factors they interact whith, they can activate or repress the splicing of an exon or intron within a certain proximity. Trans-acting factors can be roughly classified by their effect on splicing. Serine-Arginine rich proteins usually function as splicing facilitators while heterogeneous nuclear ribonucleoproteins (hnRNPs) tend to suppress splicing [Matlin et al., 2005, Wang and Burge, 2008].

Another possibility for the influence on splicing is given by the pre-mRNA itself. A certain 2D structure can either promote or inhibit binding of activators or repressors by exposing or masking the corresponding cis-regulatory elements [Warf and Berglund, 2010].

Thus, the splicing code (i.e. all influential factors in the context of splicing together with their regulation) determining the concrete composition of a mature mRNA is a highly complex mechanism depending on the combinatorial effects of the involved players, their abundance in the cell, and the competitive binding effect between activators and repressors as well as sterical inhibition by the pre-mRNA itself.

Besides the actual players involved in the splicing process, other levels of regulation exist. Post-translational modifications such as phosphorylation [Naro and Sette, 2013,

Liu et al., 2013, Zhong et al., 2009, Fackenthal and Godley, 2008] are known to play a central role in the regulation of splicing.

The coupling of splicing to transcription additionally impacts the sequence of the final mRNA. Altered RNA polymerase II elongation rates favor the inclusion of different alternative exons, while a fast elongation rate promotes skipping of exons with weak upstream 3' splice sites, slower Pol II elongation rates facilitate their inclusion [Merkhofer et al., 2014].

Moreover, epigenetic modifications can also affect and thus regulate splicing, by enhancing or preventing the accession of splicing regulatory sites [Luco et al., 2011, Brown et al., 2012, Zhou et al., 2012].

Luco et al. [Luco et al., 2011] propose an integrated alternative splice site selection model, where transcription regulators and histone modifiers alter chromatin for the recruitment of factors activating RNA Pol II elongation kinetics. Additionally, nucleosome positioning along exons as well as the enrichment of certain histone modifications might influence the recruitment of splicing factors.

While the regulation of certain splicing events or factors in a fixed scenario [Matera and Wang, 2014] as well as the mode of influence for certain levels [Luco et al., 2011] is becoming more clear due to extensive research, the general regulatory mechanisms including all relevant players and circumstances, i.e. tissue type or developmental stage, are far from being understood [Matera and Wang, 2014].

2.1.5 Alternative Splicing and Disease

Given the complexity of the process and the high number of players involved, the occurrence of errors in the splicing process is biomedically relevant, as a high number of splicing associated diseases is known. Studies suggest that about 30% of all disease causing mutations are splicing related [Xi et al., 2008]. Prominent examples are neurodegenerative diseases such as Alzheimer's [Golde et al., 1990] and Parkinson [Meshorer and Soreq, 2006] as well as cancer [Piekielko-Witkowska et al., 2010, Watson et al., 2013, Lenzken et al., 2013, Fackenthal and Godley, 2008].

The aberrant splicing events caused by mutations are often due to alterations in the splicing regulatory factors or within basic splicing signals, such as splice sites. Thus, mutations might either disrupt splicing signals located in the genome (see Figure 2.5), or impact the function of trans-acting factors.

About 10% of all pathogenic mutations are accounted to splice site alterations [Krawczak et al., 2007]. An alteration in splice-sites leads to the inclusion of introns or elongation of the sequence to splice depending on whether the mutation is located in the 5' or 3' splice site. Induction of early 3' or late 5' splice sites by mutation also affect the final sequence. Analogously, mutations in other splicing control sequences such as the branch point, enhancer or silencer sequences can provoke a reduced probability or even the total deficit of splicing.

An example, where mutations in a regulatory element lead to altered ratios of isoforms of a protein resulting in a pathologic phenotype is frontotemporal dementia and Parkinson [Fu et al., 2013, Kar et al., 2005]. Mutations in the gene MAPT were found to

be responsible for these ratio changes. The tau protein, which is encoded by this gene, is indispensable for microtubule assembly and stability. Two isoforms, one with three and one with four microtubule-binding sites, exist. The different number of binding sites is due to the inclusion of exon 10. Mutations in regulatory elements of the latter can lead to its increased inclusion, which alters the ratio of the two isoforms such that aggregation of tau is provoked [Liu and Gong, 2008].

Beyond neurodegenerative illnesses, cancer is a prominent disease class where alterations in splicing are involved. BCL-X, a gene from a familiy of apoptotic regulators, gives rise to two different isoforms with opposing functions by encoding a pro-apoptotic isoform, BCL-X_S, as well as the anti-apoptotic isoform, BCL-X_L (see Figure 2.5). The functional switch is due to an alternative 5'-splice site for one of the exons and the deregulation in the ratio of these isoforms in favor of BCL-X_L is associated to various cancer types and resistance to chemotherapy [Pajares et al., 2007, Akgul et al., 2004]. BRCA1 is a gene known to have predictive potential for breast and ovarian cancer predisposition by the genomic alterations observed. One of these known alterations is an inherited point mutation in exon 18, which disrupts an ESE and provokes skipping of the exon [Mazoyer et al., 1998].

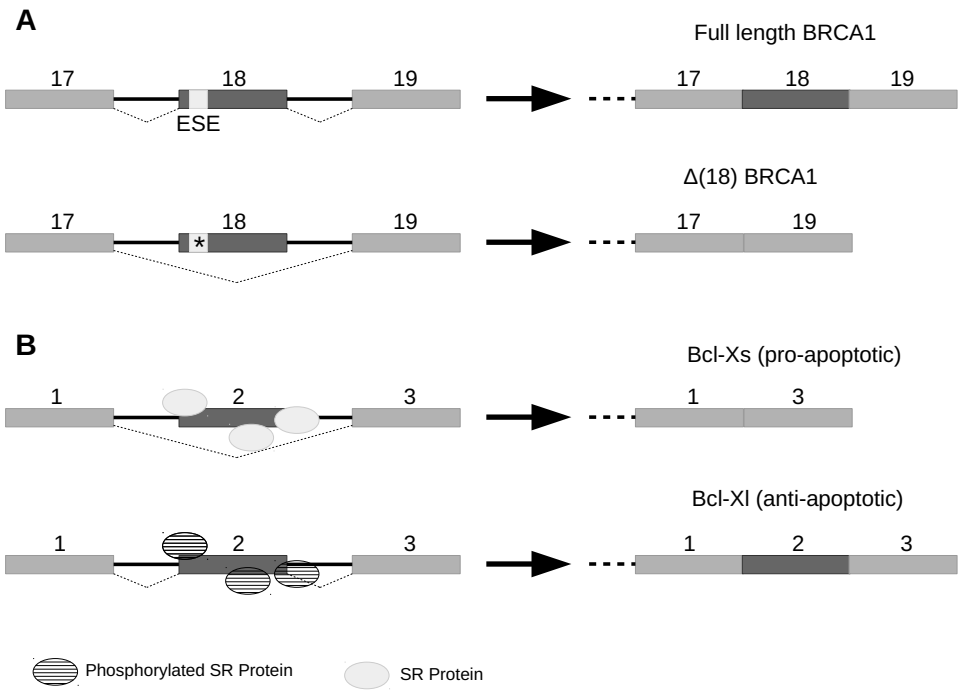


Figure 2.5: **BRCA1, BCL-X and their cancer-associated isoforms. A** A mutation (*) in an ESE located in exon 18 of BRCA1 prevents the binding of ASF/SF2 to pre-mRNA and causes inappropriate skipping of this exon. **B.** Dephosphorylation modifies the activity of SR proteins. Thereby, a splicing switch of the apoptotic regulator BCL-X_L to a pro-apoptotic isoform is induced.

The alteration of trans-acting factors by mutation can have an even higher impact on the cell. While the effects of mutations in cis-acting elements are locally restricted, splicing factors control a variety of splicing events and can therefor lead to a cascade of changes in the splicing machinery and thereby massively impact isoform types and their abundances in a cell. Non-silent mutations in the basal splicing machinery tremendously impact the whole organism and are thus usually lethal, while changes in 'secondary' splicing associated proteins might in the best case only reduce the probability for the physiologically correct splicing event. Recent publications have highlighted the oncogenic potential of aberrantly behaving splicing factors [Oltean and Bates, 2014, Sveen et al., 2015, Dvinge et al., 2016] and underlined the need for further investigations.

A trans-acting factor being subject to pathogenic alterations is, for instance, splicing factor SF2. SF2 is a proto-oncogene which is involved in several cancer-associated regulatory cascades and can lead to tumorgenesis in mice [Karni et al., 2007]. SF2 regulates alternative splicing of the oncogene Ron which influences cell motility and thus metastasis [Ghigna et al., 2005].

The role of epigenetics in the regulation of alternative splicing is being gradually elucidated [Luco et al., 2011, Iannone and Valcárcel, 2013, Zhou et al., 2014, Brown et al., 2012, Haque and Oberdoerffer, 2014]. While the concrete impact of different epigenetic modifications on the splicing process and the interplay with other splicing-associated factors remains to be determined, several links between aberrant splicing and altered epigenetic modifications in cancer have been established. Aberrant transcripts in hMLH1, for instance, are associated with lower levels of histone acetylation and specific histone methylation in gastric cancer [Zhao et al., 2017]. The tumor suppressor gene CDH1 expresses a short isoform lacking exon 8 in gastric cancer, which is attributed to different acetylation and methylation pattern of histone H3 [Li et al., 2015].

Aberrant Splicing in Lymphoma. Several publications on aberrant splicing in lymphoma exist. Besides the well known CD44 isoforms [Stauder et al., 1995, Terpe et al., 1994, Khaldoyanidi et al., 1996] cyclin D1 [Slotta-Huspenina et al., 2012, Marzec et al., 2006, Rosenwald et al., 2003], C/EBPβ [Rehm et al., 2014] and FABP7 [Lock et al., 2014] are related to aberrant isoforms in lymphoma.

For CD44, a standard isoform, CD44s, and a differing, 'variant', isoform CD44v is known. According to Khaldoyanidi et al. [Khaldoyanidi et al., 1996] expression of CD44v is elevated in peripheral blood cells of non-Hodgkin lymphoma patients and is inversely correlated with tumor progression. Additionally, response to therapy is frequently accompanied by up-regulation of CD44v. The authors thus suggest to use CD44v as a (therapeutic) marker for monitoring disease progression. Stauder et al. [Stauder et al., 1995] identify CD44v6 predominantly in aggressive lymphoma and observe a significantly shortened overall survival of the corresponding patients. Tzankov et al. investigate the role of CD44 isoforms in the subtypes of diffuse large B-cell lymphoma (DLBCL) [Tzankov et al., 2003]. The variant containing Exon 6, CD44v6, is predominantly expressed in one subgroup, the activated B cell-like types (ABC-DLBCL). Also, CD44v6 correlates with disease stage and is, for CD44 negative cases, associated with a

lower overall survival. Rosenwald et al. [Rosenwald et al., 2003] showed that mantle cell lymphoma cases mostly expressing a short Cyclin-D1 isoform, have higher Cyclin-D1 mRNA levels, a higher proliferation signature and a significantly poorer survival than cases which express the standard full-length transcript. Rehm et al. link the LAP/LAP* isoform of C/EBPβ to a lymphoma growth-promoting and -immunosuppressive environment.

2.2 High-Troughput Technologies

While most of this work is based on data derived from exon arrays, we also analyze differential splicing based on RNA sequencing data. In the following section we introduce both technologies.

2.2.1 Microarrays

Deciphering the human genome and advances in technology enabled a whole new era of research in biology and medicine. Without the knowledge of the complete human genome sequence only isolated components could be analyzed. High-throughput microarrays now allow for the simultaneous investigation of the entities of interest in a cell at a certain time point. These types of entities correspond to the different functional units of a cell. While the transcriptome, i.e. the totality of transcribed mRNAs, can be measured with expression arrays [Conway et al., 2003], the proteome of a cell is quantified with protein arrays [Zhu et al., 2001]. Another example are ChIP-chips [Buck and Lieb, 2004], where the result of a chromatin immunoprecipitation is spotted on a microarray and can thus be identified and quantified.

The basic idea of an expression microarray is to quantify the expressed mRNA of a cell by hybridization. To this end, probes are attached to the two dimensional surface of a (glass-)chip. These probes consist of short nucleotide sequences ideally representing unique sequences of the gene they target. By adding the whole transcriptome of a cell, the mRNA complementary to the gene-representing probes hybridizes to the latter and can then be quantified by the activation of an fluorescent labeling (see Figure 2.6).

A typical microarray experiment consists of several different steps. First, the mRNA of interest is extracted, and if necessary, transcribed to cDNA. After labeling, usually with a fluorescent dye, the transcriptome is added to the microarray, where hybridization of complementary sequences from probes and mRNAs is achieved. Subsequently, the dye is activated and images capture the amount of hybridized dye/mRNA. Image analysis applications detect the different spots and compute numerical quantities from the dye intensities in the picture. These numerical values, i.e. the expression values, are normalized such that technical bias between and within arrays is reduced thus facilitating the observation of biological differences. Downstream analysis can now be applied to the normalized values, including for example statistical tests for expression differences in groups, clustering algorithms for detection of groups of genes or samples, as well as classification methods for the assignment of classes.

Different types of microarrays exist. 'Two-color' microarrays are based on two channels. Thus, two biological samples, each labeled with a different dye, are hybridized to one chip. These arrays report gene expression ratios between the two channels (see also Figure 2.6). In contrast, so called single channel arrays, as, for instance, the Affymetrix Gene Chip, quantify only one biological sample and thus do not report relative expression values as illustrated in Figure 2.6.

A further distinction possibility is spotted versus in situ hybridized microarrays. While the probes on spotted microarrays are applied to the surface as a whole, the oligonucleotides on in situ hybridized arrays are spotted on the array nucleotide by nucleotide using a photolitographic process [Beier and Hoheisel, 2000]. As spotted microarrays are easier to produce and customize, they are often used as in-house designed solutions for specific questions. Their downside is a lower sensitivity compared to in situ arrays, usually produced by big companies with an highly optimized protocol. A further advantage of in situ arrays, besides their higher resolution and coverage of the genome, is their comparability across experiments.

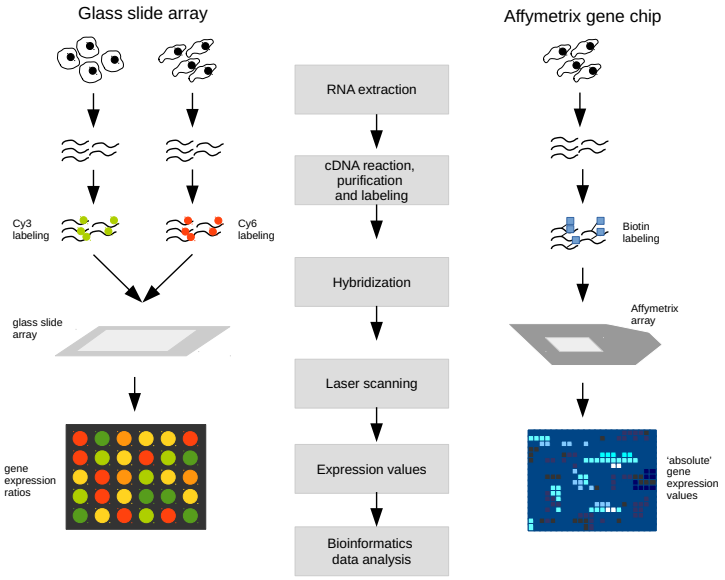


Figure 2.6: **Microarray experiment** based on dual channel (left) and single channel (right) technology. Dual channel experiments lead to relative expression values as two different samples labeled with two different dyes are quantified on one chip. A single channel microarray educes expression values based on one sample.

Irrespectively of these differences, replication is a crucial aspect for microarrays [Yang and Speed, 2002].

- Equal probes should be attached on different locations of the chip to avoid position bias and ensure a consistent signal.
- Ideally, each biological sample should be measured on different chips as a technical replicate to help identify errors of various nature in the process.

- Finally, biological replicates, i.e. samples belonging to the same condition, ideally in high numbers, are essential for the application of statistical methods and thus the deduction of reliable insights.

While microarrays are by now well established, technically mature and provide major insights in the field of biomedicine [Mischel et al., 2004, Glinsky et al., 2005], they do exhibit certain downsides. Two major drawbacks exist. First, microarrays do not quantify the exact amount of mRNA in the cell, but a function of the abundant transcripts which is dependent on several factors such as binding specificities or saturation. This allows only for reliable comparisons of a fixed gene between arrays. Second, microarrays can only measure genes that are known, as their complementary sequence has to be attached to the array in advance. Thus, unknown genes and sequences can not be represented on the microarray. In contrast, RNA sequencing does not require prior knowledge of the sequences interrogated and provides accurate transcript quantification for high coverage.

Affymetrix Exon Arrays. The *GeneChip Human Exon 1.0 ST Array* developed by Affymetrix extends the power of gene expression arrays which lack the ability to account for alternative splicing and the corresponding variety of isoform abundances. Affymetrix Exon Arrays address this issue by representing genes on their exon rather than on their gene level [Affymetrix, 2005d].

To achieve this goal, the GeneChip Human Exon 1.0 ST Array differs in several design aspects from conventional arrays, like the Affymetrix HG U133 plus 2. While gene arrays (also called 3' arrays) choose their probe location close to the 3' site of a gene trying to avoid bias induced by mRNA decay beginning at the 5' end, probes on exon arrays try to cover the whole gene evenly. Each exon is covered by on average four probes (Figure 2.7) which are later combined to one signal per exon. The higher coverage of exon arrays requires a much higher number of probes. While, for instance, the HGU 133 Plus 2 chip contains approximately 1.3 million probes, the exon array contains over 5.5 million probes (see Table 2.1) [Affymetrix, 2005d].

Addressing this equal distribution of probes, the target generation protocol for amplification has to reflect probe locations. In contrast to most 3' arrays, which use primers targeting the poly-A tail of mRNAs, the GeneChip Whole Transcript (WT) Sense Target Labeling Assay [Affymetrix, 2006] applied for the exon array uses randomly attaching primers. Thus, a bias towards the 3' end of transcripts is avoided and the technique of using random primers leads to evenly distributed probes (see Figure 2.7).

Probe types. Apart from confirmed exons, the exon array explicitly also contains uncertain and predicted exons to cover as many exons as possible. All probes on the array are divided into three categories according to their reliability [Affymetrix, 2005d]:

1. The first category comprises the *core probes*. These probes are derived from RefSeq [Pruitt and Maglott, 2001] transcripts or full-length mRNAs.

2. The *extended* loci contain all core probes as well as all cDNA based loci. Among these are ESTs, mRNAs from Genbank which are not annotated as being full-length, as well as microRNA annotations [Affymetrix, 2005b].

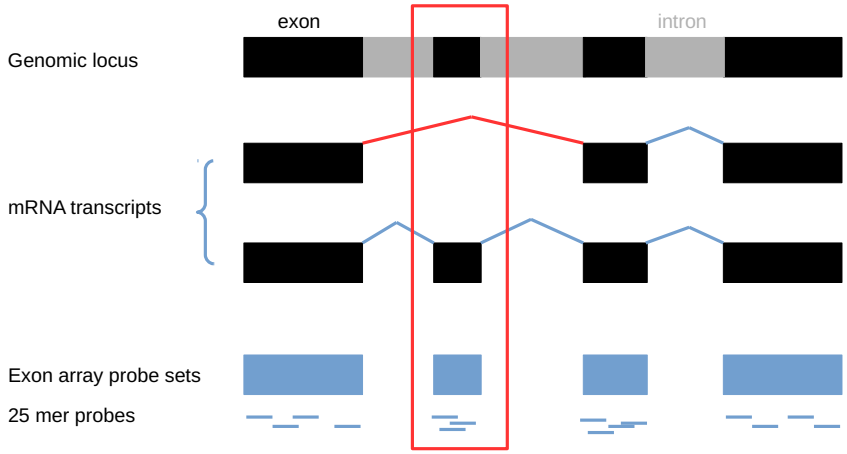


Figure 2.7: **Exon array probe coverage** Evenly distributed probes across all exons enable the detection of the skipped exon in the first transcript.

	GeneChip HG U133 Plus 2.0	GeneChip Human Exon 1.0 ST Array
Probes per gene	11-20	~40
Probes per array	~1,300,000	~5,500,000
Probe sets per array	54,000	1,400,000
Background probes per array	650,000	40,000

Table 2.1: **Comparison of 3' arrays and exon arrays** based on different attribute numbers.

3. The category *full* loci encompasses all probes from the extended loci plus loci derived from ab-initio gene predictions.

To avoid cross hybridization, sequences of all probes have been compared to each other. Sequence similarities to untranslated regions are ignored, i.e. not excluded from the set of probes, to avoid unnecessary rejection of thermodynamically favorable probes. Affymetrix classifies all probe sets into three categories according to their cross-hybridization potential [Affymetrix, 2005d]:

- The about 1.25 million unique probe sets contain only probes that have no known potential for cross-hybridization with probes of other probe sets.
- Approximately 70,000 probe sets are categorized as similar. These sets contain probes that are candidates for cross-hybridization, but all probes of the respective probe set interrogate the same genomic region, i.e. the same gene.
- Approximately 200,000 mixed probe sets exhibit inconsistent hybridizations, i.e., they might hybridize to different locations in the genome.

Background Probes. To estimate a reliable background signal, Affymetrix 3' arrays contain as many perfect match probes as mismatch probes [Affymetrix, 2005d]. The mismatch probes differ in one base from the corresponding perfect match probe and are used to calculate the unspecific binding signal strength. For exon arrays, a much higher number of probes is required to cover all exons. Accordingly, these arrays do not contain probe-specific background probes, but only a set of about 40,000 background probes in total. To account for the effects of GC-richness on hybridization strengths, background probes are binned according to their GC-content. 26 bins of different GC contents are defined, each containing approximately 1,000 background probes. For each bin, a separate null distribution for probes with this GC content is calculated and used to estimate robust confidence values. Furthermore, these 40,000 background probes are divided into genomic and antigenomic background probes, and either can be used to estimate a background signal. Binning differs according to whether a probe is of genomic or antigenomic origin:

- *Genomic* background probes match to regions that are not likely to be transcribed. To produce reasonable background probes mismatches have been introduced. Each bin of GC content is covered by about 1,000 mismatch probes.

- *Antigenomic* background probes originate from sequences that are not found in the human, mouse or rat genome. Therefore, they are not expected to cross-hybridize with transcribed human DNA. The 26 bins range from a total absence of G and C nucleotides to a GC content of 100%.

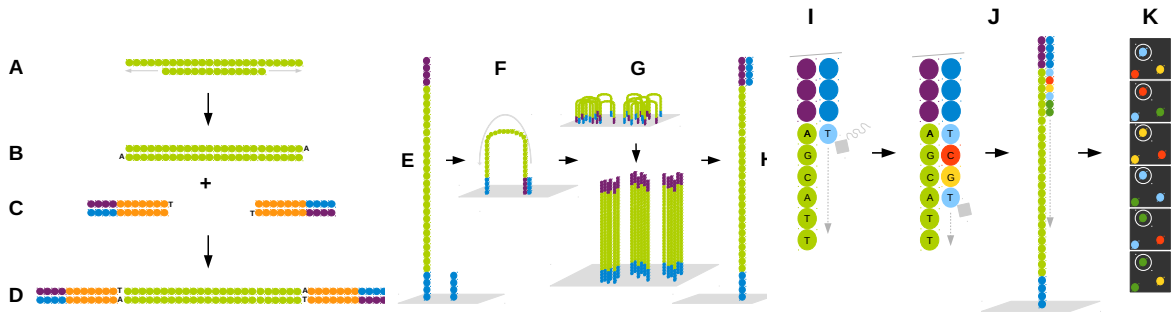
2.2.2 High-Throughput Sequencing

High-throughput sequencing (HTS), a more recent technology in the field of molecular genetics compared to microarrays, has now been widely used for over a decade and provides several advantages over microarrays. HTS is not only able to capture sequences not previously known, as they do not depend on pre-designed probes, but they also have a much higher dynamic range, than microarrays [Nagalakshmi et al., 2008]. While for microarrays expression measurement is limited by signal saturation at the high and background at the low end, RNA sequencing provides discrete read counts. This allows for a better qualitative and quantitative transcriptome acquisition. The importance of this fact is highlighted in studies stating that about three quarters of the genome are probably transcribed [Djebali et al., 2012].

HTS allows for a whole new perspective on the genome. While microarrays provide solutions for most investigation levels such as chromatin immunoprecipitation or single nucleotide polymorphisms (SNPs), sequencing based solutions are more accurate and sometimes even enable quantification of events or entities for the first time. SNP arrays, for instance, exist, but they can only be used for predefined mutations on fixed positions. Thus, only a tiny fraction of all possible mutations can be detected. As an array can never represent the whole genome or transcriptome with all its combinatorial products, it is the technology of HTS, that allows for the detection of all novel splice isoforms.

Sequencing also opens insights into the wide field of RNA classes not formerly captured. These transcribed entities such as small nuclear RNAs (snoRNAs) or long non-coding RNAs (lncRNAs) have moved into the focus of attention only since HTS made them accessible [Qureshi and Mehler, 2012].

Several prevailing technologies use the sequencing-by-synthesis approach, such as Illumina. For an exemplary workflow we outline the basic proceedings of HTS in Figure 2.8.



(a) Library preparation (b) Cluster generation (c) Sequencing

Figure 2.8: **Overview on Illumina sequencing.** For *library preparation (a)* the DNA is fragmented (A), ends are repaired (B), adapters are ligated (C) and ligated DNA is selected (D). In a next step, *clusters are generated (b)*: DNA is attached to a flow cell (E), bridge amplification is performed (F), clusters are generated (G) and sequencing primers are annealed (H). For *sequencing (c)*, the first base is extended, read and deblocked (I). This step is repeated until the strand is fully extended (J) and base calls can be generated (K).

Library Preparation. The initial set of RNA is gathered in a library. RNA is extracted from the cells of interest and further selected depending on the aim of the study. If, for instance, mRNA is the target of investigation, an enrichment of poly-A carrying entities might be performed. miRNAs, on the other hand, can be selected by size filtering. The resulting set of targets is subsequently fragmented into smaller pieces and reverse transcribed to cDNA. The resulting double-stranded cDNA fragments are ligated with adapter sequences on both ends and are thus ready for amplification.

Cluster generation. The small double-stranded, adapter ligated cDNA fragments resulting from the library preparation are homodimerized and attached to the surface of a so-called flow cell, a glass slide with eight flow channels. Subsequently, a bridge amplification is performed. The name refers to the bridge-like structures arising from the cDNA fragments bound to the glass slide on both ends due to the adapters attached. These fragments are amplified by synthesization of the complementary strand. The double stranded structures are denaturized and the cycle repeats. This procedure leads to many copies of one fragment in close spatial proximity, referred to as local clusters [Reuter et al., 2015].

Sequencing. Using the agglomerations of identical fragment representatives, sequencing by synthesis is performed. Each sequencing cycle takes four steps according

to the four nucleotides. The nucleotides are labeled by type and can thus be detected when they are incorporated in the sequence. Therefor, laser-excited light is used which is captured by a camera after each step. From these images and the corresponding light, the sequence structure can be reconstructed.

Sequencing is being constantly improved in accuracy and cost, and the initial goal to reduce prices per genome to 1,000 US Dollars has been practically achieved [Hayden, 2014]. Yet, the technology still suffers from shortcomings, as it is for example difficult to optimize accuracy, speed and cost at the same time. While for instance Illumina indicates its average error rate with 1%, other solutions exist, which show a much higher error rate of 5%-40% with the advantage of a significant speedup [Goodwin et al., 2015].

For transcriptomics, short reads state a severe problem for encountering spliced isoforms, especially in concrete isoform inference, as they rarely span multiple splice junctions. However, not only technological problems impose challenges. The probably biggest challenge is to develop adequate computational analysis routines for the vast amounts of data produced by HTS. Though several efforts exist [Nekrutenko and Taylor, 2012], for most omics-technologies a standard is not yet established. With the lack of uniform analysis routines it is difficult, to compare technologies and thus assess their accuracy.

Data Analysis. Each investigation level and each corresponding technology have their specific requirements and specificities concerning analysis. In this work we are focusing on transcriptome sequencing, especially mRNA sequencing. The basic steps necessary to gain insights from the data produced are the alignment of the reads, expression quantification, detection of differentially expressed entities, and, optionally, downstream analyses.

1. **Alignment.** The alignment of reads, also referred to as read mapping [Li and Homer, 2010], searches for the matching position of a read in the genome and associates this position with the read. One of the widest used algorithms is implemented in the Burrows-Wheeler aligner [Li and Durbin, 2009] though many others exist [Li and Homer, 2010].

2. **Expression Quantification.** The information gained in the alignment step allows for the quantification of expression for the entities of interest. Therefor, all reads falling in the region of such an entity are summed up to obtain an absolute expression level.

3. **Differential Expression.** With the computed count values, differences in expression between two or more conditions can be assessed. Depending on the aggregation level of the data, differences in gene expression, transcript expression or even on single exon expression can be detected.

4. **Downstream Analysis.** Further analyses such as classification, clustering, enrichment analysis or network analysis can be applied to identify biological subclasses, infer functional associations for groups of entities or determine biomarkers.

2.3 Data

Although the focus of this work lies on methodological aspects, the ultimate aim of biomedical data analysis is an information gain to understand biological conditions. Throughout this work, we will apply the methods to expression data from different lymphoma samples pursuing the aim of elucidation of aberrant splicing events as well as their regulation in different lymphoma subtypes.

2.3.1 Lymphoma Subtypes

Lymphoma is a cancer based on lymphatic cells. The lymph system is part of the immune system and comprises different organs such as lymph nodes, spleen, bone marrow, thymus and tonsils (see Figure 2.9). As of its immune system based origin, lymphoma cells can be of T-cell or B-cell origin. Two main classes of lymphoma exist. While Hodgkin lymphoma are usually curable, non-Hodgkin lymphoma form the therapeutically more demanding class [Shankland et al., 2012]. The set of expression data used in this study consists solely of non-Hodgkin lymphoma which are subdivided in several subgroups with different cellular origin.

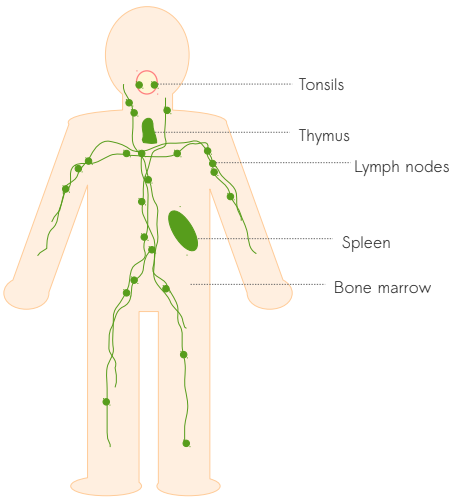


Figure 2.9: **Anatomy of the lymph system.** The lymph system comprises the lymph vessels and lymph organs which include lymph nodes, tonsils, thymus, spleen, and bone marrow.

- **Diffuse large B-cell lymphoma (DLBCL).** Diagnosis is based on several aspects of the disease such as cellular morphology, immunohistochemistry, and gene expression. While the first two methodologies are beyond the scope of this work, gene expression studies on DLBCL reveal a distinction between two groups, germinal center B-cell-like (GCB) and activated B-cell-like (ABC). Cases of GCB are usually associated with a rather positive prognosis, while ABC cases are expected to have a poor outcome.

- **Mantle cell lymphoma (MCL).** MCL is a rather rare B-cell lymphoma with a characteristic translocation at t(11;14)(q13;q32) leading to an over-expression of cyclin D1 [Campo et al., 1999], a cell cycle gene contributing to the abnormal proliferation of the malignant cells.
- **Follicular lymphoma (FL).** FL, another B-cell lymphoma, is usually characterized by a translocation between chromosome 14 and 18 (t(14;18)(q32;q21)). This alteration results in the over-expression of BCL-2, a gene with anti-apoptotic function. This mutation drastically reduced susceptibility to apoptosis, a key feature of malignant cells.
- **Anaplastic large-cell lymphoma (ALCL).** This T-cell based lymphoma subtype also exhibits a typical translocation (t(2;5)(p23;q35)), resulting in the upregulation of a tyrosine kinase with oncogenic properties.
- **Peripheral T-cell lymphoma (PTCL).** For this T-cell based lymphoma subtype, no prominent translocation is known.
- **B-cell chronic lymphocytic leukemia (CLL)** Several chromosomal deletions (del 11q, del 13q, del 17p) as well as a trisomy in chromosome 12 characterize this B-cell based disease.

2.3.2 Lymphoma Data

In this work, we will use data generated by both technologies described in Section 2.2, exon arrays as well as RNA sequencing data. The majority of the data is based on exon arrays, while a subset of the biological samples represented on exon arrays has also been sequenced.

Lymphoma subtype	No. of exon array samples	No. of RNA sequencing samples
BL	3	-
CLL	14	-
DLBCL	40	3
ALCL	10	-
PTCL	6	-
MCL	12	-
FL	22	-
Tonsil	9	3

Table 2.2: **Sample number per lymphoma subtype** based on exon array and RNA sequencing expression quantification.

In total, 116 exon arrays were used to measure mRNA expression of 116 biological samples from seven lymphoma subtypes as well as a control group of tonsil samples. Group sizes vary drastically. While the smallest group, Burkit lymphoma, comprises

only 3 samples, the largest group, diffuse large B-cell lymphoma consists of 40 samples. For an overview see Table 2.2.

The subset of biological samples which were also subject to sequencing consists of three samples for DLBCL as well as three samples from the tonsil control group. As the biological source is identical, this procedure provides an interesting possibility for comparison of both technologies. DLBCL are a very heterogeneous disease, where several subtypes are known. Thus, research can benefit a lot from a more detailed characterization.

3 Detection of Differential Splicing

The analysis of differential splicing (DS) is crucial for understanding pathophysiological processes in cells and organs [Srebrow and Kornblihtt, 2006]. A widely used technique for studying DS are exon arrays (see Chapter 2.2.1). Over the last decade, a variety of algorithms detecting DS events from exon arrays has been developed. However, no comprehensive, comparative evaluation including assessment of the most important data features has been conducted so far. To this end, we created multiple data sets based on simulated data to assess strengths and weaknesses of several published methods as well as a newly developed method, KLAS. Additionally, we evaluated all methods on two cancer data sets that comprised RT-PCR validated results.

3.1 Aim

The detection of altered expression on the exon level is a challenging task. It is by far more demanding than gene-based analyses since relative changes in exon expression levels might be more subtle, which makes it harder to distinguish signal from noise.

Furthermore, changes in the expression of the gene containing the exon have to be taken into account to avoid false positives as well as false negatives. To accomplish this task, exon expression is usually normalized to the corresponding gene expression. Figure 3.1 visualizes a situation in which the second exon of tissue A is differentially spliced, as it is included less in the final transcript than in tissue B. Yet, as gene expression is higher in tissue A, a comparison based on exon expression only, would lead to the lowest evidence for DS in the second exon. Thus, normalization of exon expression with the corresponding gene expression intensity levels gene expression differences between groups compared and enables the detection of different exon in- or exclusion levels, i.e. differential splicing.

Besides these general difficulties, various parameters impact the correct prediction of DS. The expression level, already impairing the detection of differential gene expression [Draghici et al., 2006], imposes an even higher difficulty in the discovery of exon expression alterations. The number of exons in a gene is also known to influence performance, specially for statistical methods [Rasche and Herwig, 2010]. Splicing events might not necessarily occur in all samples of a condition, which could have a different effect on method performance. Another impacting factor could be the number of samples in the conditions under examination.

The wealth of existing expression data from exon arrays constitutes an excellent basis for many scientific questions. This led to a variety of algorithms for differential splicing detection developed over time. Different approaches were taken to solve the task.

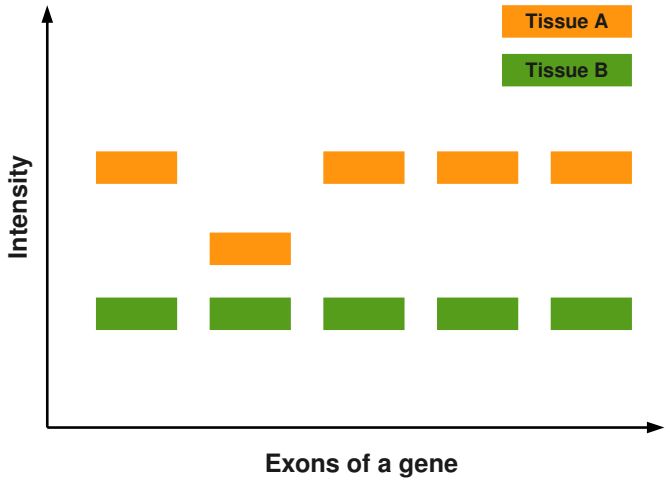


Figure 3.1: **Differential exon expression.** The second left exon in tissue A is differentially spliced. A comparison on exon level only would lead to the opposite of the desired result, as the only exon differentially spliced would gain the lowest evidence for DS since the expression of this exon is the most similar in tissue A and B.

Most of the methods, such as MIDAS [Affymetrix, 2005a], use a statistical approach. Other methods combine statistics with the exploitation of the preprocessing results (e.g. FIRMA [Purdom et al., 2008]). SplicingCompass [Aschoff et al., 2013], a graphical approach, is based on angles between exon expression vectors, while ARH, relies on information theory [Rasche and Herwig, 2010]. These large methodological differences make it impossible to compare methods analytically, which calls for careful empirical studies to identify the best tool for a given scenario. We thus conducted the - to our knowledge - most comprehensive comparative assessment of algorithms for DS detection on exon arrays. We compared and evaluated nine different methods for the detection of differential splicing from exon arrays. Subsequently, we discerned the performance for each method over a range of different parameters. Using a comprehensive artificial dataset, we compared the impact of different expression levels, numbers of exons per gene, different amounts of differentially spliced samples per condition as well as the influence of different group sizes. Additionally, we applied all methods to two well studied and partly RT-PCR validated cancer data sets [Gardina et al., 2006, Langer et al., 2010].

3.2 Methods for Differential Splicing Detection

In the following, all methods evaluated in this work are introduced in detail. We included, to our knowledge, all published methods until 2014 for which an implementation was available: MADS, MIDAS, SI [Affymetrix, 2005a], PAC [Affymetrix, 2005a, French

et al., 2007], ANOSVA, ARH, SplicingCompass [Aschoff et al., 2013] and FIRMA. Furthermore, we incorporated the novel method KLAS [Jentsch, 2011]. Note that we did not use FIRMA for evaluation on artificial data, as we used the model proposed by the authors of [Purdom et al., 2008] on the basis of which FIRMA was developed for the generation of our data. However, we applied FIRMA to the two experimental data sets. We had to leave out methods with no implementation available, like Remas [Zheng et al., 2009].

The notation defined here refers to the methods presented in the subsequent chapters. Throughout this chapter, the comparison of two conditions $h \in c, t$ is assumed, a *control* c as well as a *treatment* group t. The groups shall consist of $s = 1 \dots q$ samples in the control group and $s = 1 \dots r$ samples in the treatment group. While some methods are in theory applicable to more than two conditions, we limit ourselves to the application of two for comparison purposes. The two case scenario is by far the most common application.

The probes p on the chip measure a certain intensity λ which depends on condition h, gene g (comprising n exons), exon e, sample s and probe p:

$$\lambda_{g,e,p,h,s} \in \mathbb{R}^+ \tag{3.1}$$

A gene denoted by g shall contain $e = 1 \dots n$ exons. The expression of a gene g or exon e is denoted by Ψ or ψ respectively which is an aggregate f over the exons for g or probes of an exon for e. More precisely,

$$\Psi = f(\lambda_{e,h,s}) \in \mathbb{R}^+ \tag{3.2}$$

$$\psi = f(\lambda_{p,h,s}) \in \mathbb{R}^+ \tag{3.3}$$

ANOSVA

ANalysis Of Splice VAriation (ANOSVA) [Cline et al., 2005] uses a two way ANOVA to detect potential alternative splicing events. Hereby, each observation, i.e., each probe intensity is described in dependence of a factor for exon dependence α_e as well as a factor for the biological condition β_h. More precisely, the following linear model is used to fit the observed data:

$$log2(\lambda_{h,e,p,s}) = \mu + \alpha_e + \beta_h + \gamma_{e,h} + \epsilon \tag{3.4}$$

The intensity value $log2(\lambda_{h,e,p,s})$ denotes the observed $log2$ expression intensity of probe p in exon e of sample s in condition h. The term μ denotes an intercept used for all probes while ϵ denominates an error term. Alternative splicing is detected by use of the term $\gamma_{e,h}$. It describes the interaction effect for all combinations of α_e and β_h, i.e. additional signal not captured by the two factors on their own. Thus, non-zero interaction terms are thought of as evidence for differential exon inclusion.

ARH

Alternative splicing robust prediction based on entropy (ARH) [Rasche and Herwig, 2010] is, unlike other approaches, not based on correlation or statistical tests, but applies an information theoretic approach based on Shannon's entropy. An advantage of this method is that it overcomes problems like the dependency of a score on the number of exons or the inherent variability in exon expression intensities. A drawback of ARH is, that it can be applied only in the case of two different conditions.

First, the exon splicing deviation $\delta_{g,e}$ between the two conditions is computed by subtracting the median $log2$ ratio of the exon expressions from all median $log2$ ratio exon expressions of the associated gene.

$$\delta_{g,e} = log2\left(\frac{\psi_{g,e,t}}{\psi_{g,e,c}}\right) - median_{e=1,\ldots,n}\left(log2\left(\frac{\psi_{g,e,t}}{\psi_{g,e,c}}\right)\right) \tag{3.5}$$

The absolute value of the splicing deviation is turned into the probability $p_{g,e}$ of exon e being differentially spliced.

$$p_{g,e} = \frac{2^{|\delta_{g,e}|}}{\sum_{e=1}^{n} 2^{|\delta_{g,e}|}} \tag{3.6}$$

Next, an entropy is computed for each gene, indicating whether the splicing probabilities are equally distributed.

$$H_g(p_{g,1},\ldots,p_{g,n}) = -\sum_{e=1}^{n} p_{g,e} \cdot log2(p_{g,e}) \tag{3.7}$$

The theoretical maximum $max(H_g)$ of the entropy is $log2(n)$. To make the entropy independent of the number of exons, it is subtracted from the theoretical maximum.

$$max(H_g) - H_g = log2(n) - H_g(p_{g,1},\ldots,p_{g,n}) \tag{3.8}$$

Now the final score ARH for gene g can be computed to indicate whether g is alternatively spliced.

$$ARH_g = \frac{Q_{75,g}}{Q_{25,g}} \cdot (max(H_g) - H_g) \tag{3.9}$$

The weighting factor $\frac{Q_{75}}{Q_{25}}$ accounts for the strength of deviation within g. Here, Qxx denotes the interquartile range of expression range of the $xxth$ quartile. By the computation of a background distribution for ARH values from various datasets, values larger than 0.03 are considered to be an indication for alternative splicing by the authors.

FIRMA

Finding Isoforms using Robust Multichip Analysis (FIRMA) [Purdom et al., 2008] has the major advantage that it can also be used in the scenario of no predefined groups

or if alternative splicing events are not consistent in the given conditions. The basic idea is to use a fitted linear model for expression estimation and deduce a score for alternative splicing for each exon from the model parameters. Robust Multichip Analysis (RMA) is used to determine these parameters. The fitting of the linear model for expression estimates leads to the possibility to compute the difference between estimated and measured expression. This difference is taken as basis for the computation of a score for differential exon expression. By doing so, the problem of alternative splicing detection essentially is converted to one of outlier detection.

The model fitted by RMA for every gene g contains a chip effect term a_s for the sth chip, a probe effect term b_p for the pth probe and an error term $\epsilon_{s,p}$ An estimate for the background corrected and normalized expression level of a gene is computed as follows:

$$log2(\lambda_{p,s}) = a_s + b_p + \epsilon_{s,p} \tag{3.10}$$

For exon arrays the model can be adjusted by introducing o_e - the relative change in exon expression for exon e - and the interaction $\delta_{s,e}$ between chip and exon as well as a new error term $\epsilon_{s,e,p}$.

$$log2(\lambda_{e,p,h,s}) = a_s + o_e + \delta_{s,e} + p + \epsilon_{s,e,p} \tag{3.11}$$

As the parameter $\delta_{s,e}$ represents the difference between an exon in sample s and the expected expression for this exon, the parameter can be seen as a measure of differential splicing. Instead of fitting an exon level model, the gene level model is fitted with the exon array data to improve robustness. $\delta_{s,e}$ is then estimated by using the residuals $l_{s,p,e}$ of the fitted gene level model.

$$l_{s,p,e} = y_{s,p,e} - \hat{a}_s - \hat{b}_p \tag{3.12}$$

The actual value is computed by averaging over all residuals of an exon. The final score for alternative splicing is now calculated as:

$$F_{s,e} = median_{p \in e} \left(\frac{l_{s,p,e}}{\sigma} \right) \tag{3.13}$$

where σ, an estimate of the standard error, is derived from the median absolute deviation (MAD) of the residuals. By introducing this parameter the score is made more comparable between different genes.

KLAS

KLAS, Kullback-Leibler alternative splicing [Jentsch, 2011] uses a similar approach as ARH, but relies on the Kullback-Leibler divergence in the last step. The Kullback-Leibler divergence is an indicator for the variety of two probability distributions. For

each condition $h \in t, c$ the deviation λ of the expression of every exon e from its gene g is computed.

$$\lambda_{e,t} = \psi_{e,t} - \Psi_t \tag{3.14}$$

$$\lambda_{e,c} = \psi_{e,c} - \Psi_c \tag{3.15}$$

Subsequently, these deviations are turned into a probability distribution per gene and condition, such that the contribution of every exon to the expression of the gene can be denoted by

$$p_{g,e} = \frac{2^{\delta_{e,h}}}{\sum_{e=1}^{n} 2^{\delta_{e,h}}} \tag{3.16}$$

This is a major difference to ARH, which assesses one probability distribution for both conditions based on the deviation from the median exon ratio between conditions. To account for the deviation within a gene, the interquartile range is computed.

$$Q_t = \frac{quant_{0.75}(\delta_{e,t})}{quant_{0.25}(\delta_{e,t})} \tag{3.17}$$

$$Q_c = \frac{quant_{0.75}(\delta_{e,c})}{quant_{0.25}(\delta_{e,c})} \tag{3.18}$$

This step is similar to ARH, yet here it is used to compare two conditions based on a modified Kullback-Leibler divergence as formulated in Equation 3.19 instead of the Entropy corrected by its theoretical maximum as for ARH.

$$KLAS(g) = Q_t \cdot \sum_{e=1}^{n} p_{e,t} \cdot log\left(\frac{p_{e,t}}{p_{e,c}}\right) + Q_c \cdot \sum_{e=1}^{n} p_{e,c} \cdot log\left(\frac{p_{e,c}}{p_{e,t}}\right) \tag{3.19}$$

The main difference between KLAS and ARH is thus, the level at which the entropy, respectively the Kullback-Leibler divergence, (i.e. relative Entropy), is computed. While entropy is a feature of one probability distribution, the Kullback-Leibler divergence is an indicator for the variety of two probability distributions. Where ARH is constrained to case control studies, the approach to establish the probability distribution within the samples allows extension of the analyses to more than two conditions.

MADS

Microarray analysis of differential splicing (MADS) [Xing et al., 2008] consists of three steps: background correction, summarization and detection of differential splicing events. For background correction, a sequence-specific linear model with many parameters is fit to predict the background intensity for each probe. The predicted background intensity is then subtracted from the observed signal. Genomic as well as antigenomic background

probes are used to train the model. The advantage of this background model is a nucleotide and position specific model for the 25mer probes.

In the second step the probes with the highest correlation over all samples are selected for each gene by application of hierarchical clustering. The Li-Wong [Li and Wong, 2001] model is fitted to these probes to compute an estimate of gene expression. Only probes are kept that show high correlation between the background corrected values and the corresponding gene signal estimates over all samples. Similar to iterPLIER [Affymetrix, 2005c], this procedure is repeated until the number of probes stabilizes.

To determine differential splicing, first the Splicing Index is calculated for every probe.

$$SI_{MADS} = \frac{\lambda_{e,p,h,s}}{\Psi_{h,g}} \tag{3.20}$$

A t-test is applied to determine the significance of the calculated Splicing Indices. The resulting probe level p-values are transformed.

$$x = -2log(p - value) \tag{3.21}$$

They are now following a χ^2_2 distribution under the assumption of no differential splicing. The sum of these transformed p-values follows a χ^2_{2k} distribution with k indicating the number of probes. Using this sum, an exon-level p-value is calculated and all exons are ranked according to it. The final results are filtered for potentially cross hybridizing probes. For applicability reasons we adopted MADS to work on our synthetic data, i.e. we did not apply a MADS-specific background correction and modified MADS to work on the exon level. We will therefore refer to this modified method as MADS'.

Midas

The idea behind Microarray Detection of Alternative Splicing (MIDAS) [Affymetrix, 2005a] is similar to the idea of SI and is followed by a statistical test. Like for other DS detection methods, it is assumed that if the ratio of the expression level of an exon to the expression level of the corresponding gene is constant over all samples, no differential exon expression has occurred. Thus, for all exons, the exon-gene ratio, i.e. a normalized intensity NI is computed per sample.

$$NI = \frac{\psi_{e,d}}{\Psi_d} \tag{3.22}$$

Their variance is tested for statistical significance using, for instance, ANOVA in case of multiple group comparisons. In the two condition scenario a t-test is applied.

PAC

The underlying assumption in Pattern based correlation (PAC) is the proportionality of exon expression to its corresponding gene expression. Deviation from exon to gene expression results in low correlation and therefore indicates DS [French et al., 2007,

Affymetrix, 2005a]. To identify such potential splicing candidates, PAC computes an expected expression level per exon and compares this value to the actually measured expression of the exon.

The estimated expression of exon e is defined as

$$estimate(\psi_{g,e,s}) = \Psi_{g,s} \cdot \frac{mean(\psi_{g,e,s})}{mean(\Psi_{g,s=1,...,q+r})} \tag{3.23}$$

Scaling of the gene expression of gene g in sample s with the ratio of the average exon expression $mean(\psi_{g,e,s})$ of exon e over all samples $1...q+r$ to the average of all gene expressions $mean(\Psi_{g,s=1,...,q+r})$ of gene g in all samples $1...q+r$ leads to the final expression estimate $estimate(\psi_{g,e,s})$.

This estimate is compared to the actually measured exon expression $\psi_{g,e,s}$ and is expected to be the same in the case of no alternative splicing. Originally, PAC was developped as a multi-condition method, where the estimated and measured exon expression can be correlated. A low correlation is an indication for alternative splicing. In the two group scenario correlation is not applicable as correlation will either be 1 or -1 except for the case of exact equality. The difference between the two is assessed instead.

$$estimate(\psi_{g,e,s}) - \psi_{g,e,s} = 0 \tag{3.24}$$

Splicing Index

The splicing index [Affymetrix, 2005a, Srinivasan et al., 2005, Clark et al., 2002] can be regarded as the exon level analog of the gene level fold change. It is a measure for the difference of exon specific expression between samples. To avoid false positives due to differing gene expression between conditions, each exon is set into relation to the gene expression of the gene it originates from.

As a first step, the expression level of each exon e in sample s is normalized to its corresponding gene expression g computing a normalized intensity NI.

$$NI_e = \frac{\psi_{e,s}}{\Psi_s} \tag{3.25}$$

Based on these normalized measures, the Splicing Index $SI(e)$ measures the relative change between two conditions:

$$SI(e) = \frac{NI_{e,t}}{NI_{e,c}} \tag{3.26}$$

In the case of sets of samples, i.e. for multiple samples per condition, several procedures are applicable. On the one hand, the ratio of the medians of each condition can be calculated. In the case of paired samples, one can also compute the SI for all pairs from the two conditions, and take the median as an indicator of alternative splicing. A more sophisticated approach tests for differences in group specific SIs by applying a statistical test. The SI can only be applied in a two-condition scenario.

Splicing Compass

We adopted Splicing-Compass, originally developed for NGS data [Aschoff et al., 2013], to also work for exon array data. The idea is to access the significance of difference between angles spanned by exon vectors in one condition compared to the ones in the other condition.

Let v_g be the gene expression vector for gene g, where every dimension of g stands for an exon expression value of an exons e contained in g.

$$v_g = (\psi_1, \ldots, \psi_i, \ldots, \psi_n) \in \mathbb{R}^n \tag{3.27}$$

If the isoform ratios of a gene are constant throughout the conditions, i.e. no alternative splicing event occurs, the expression values of this gene will be approximately parallel between conditions. This leads to a small angle between the two, even in the case of differential isoform expression between conditions.

To test for differences in splicing, all pairwise angles $\binom{q+r}{2}$ are computed as follows.

$$\Phi_{g_{n_i}, g_{n_j}} = \arccos\left(\frac{v_{g_{n_i}} \cdots v_{g_{n_j}}}{\left\|v_{g_{n_i}}\right\| \cdots \left\|v_{g_{n_j}}\right\|}\right) \tag{3.28}$$

Subsequently, a one-sided t-test is applied to determine whether splicing angles within conditions are significantly smaller than splicing angles between conditions. To account for multiple testing errors, a Benjamini-Hochberg correction is applied.

3.3 Data

Most studies comparing the performance of methods for DS detection are based on either synthetic or real data sets. Both approaches have their strong points, while they do not lack certain downsides. Synthetic data allows for a controlled parameter setting, enabling an easier deduction of reasons for success or failure of methods applied. On the other hand, simulation models only approach real scenarios and can thus be subject to discussion. Real data is limited to specific scenarios, as evaluation is laborious and expensive. Nevertheless, it is ultimately the latter that analysis methods have to work on. Thus, we decided to include both evaluation strategies in our study.

Synthetic Data

The performance of each method for differential splicing detection is influenced by many factors. A systematic analysis of the properties inherent to the different methods can only be achieved by using specifically designed artificial test data. To this end, we generated a range of synthetic data sets using the model from [Purdom et al., 2008]:

$$y_{i,j} = log2(B_j + I_{i,j} \times 2^{p_i + c_j}) + \epsilon_{i,j} \tag{3.29}$$

where $y_{i,j}$ equals the $log2(PM)$, with PM denoting the perfect match probe, for chip i and probe j. B_j denotes the additive background modeled as a log-normal variable. The chip effect c_i is normally distributed while the probe affinity p_j and the residuals $\epsilon_{i,j}$ are modeled as mean-zero normal variables. To control probe expression, the indicator variable $I_{i,j}$ is set to 1 in the case of presence and to 0 otherwise. According to the authors, values for simulation parameters were chosen by estimates of representative values from real data. We decided to use this model, as it is the most fine-grained we were aware of.

We applied multiple parameter allocations in many combinations (see Table 3.1) using the default settings of the model. Specifically, we studied the influence of the number of exons per gene ($enum \in \{10, 30\}$), the expression intensity ($expr \in \{high, low\}$), the number of samples ($snum \in \{15:15, 15:5\}$) per group as well as the percentage of differentially spliced samples ($pcnt \in \{60, 100\}$). The combination of these four parameters with two allocations each led to a total of 16 scenarios yielding a detailed insight that is important when choosing the adequate method for a given dataset or for a certain purpose. For modeling of expression intensity, we used the proposed model parameter values of $cmean = 7$ in the low and $cmean = 10$ in the case of high expression [Purdom et al., 2008].

Parameter	**short**	**value 1**	**value 2**
samples per group	*snum*	15 vs. 5	15 vs. 15
exons per gene	*enum*	10	30
expression intensity	*expr*	high	low
percent differentially spliced samples per group	*pcnt*	60 %	100 %

Table 3.1: **Values used for the different parameters tested.** The combination of 4 parameters with two possible values leads to 16 data scenarios.

In each scenario we generated 200 simulated genes. While 100 genes were specific to the parameter criteria in addition to displaying differential splicing events (true positives (TP)) the remaining 100 genes, designed as true negatives (TN), show no altered exon expression. Thus, probably the most challenging of the 16 data sets for a DS detection method (see also Table 3.1) consisted of (1) one condition containing 15 samples and a second condition containing only 5, (2) low expression intensity, (3) only 60% of the samples in a group exhibiting differential splicing and (4) a high number of exons per gene.

It is undoubtedly more demanding to detect DS in a small group where not all samples display the event than in a large group under the same condition. Concerning the scenarios with an imbalance in group size, we therefore switched the DS event containing group for half of the TP genes. Thus, in settings with one condition containing 15 samples, the other one 5 samples and DS was only simulated in 60% of the samples,

half of the TP genes show the DS event in the small group and half of them in the large group.

Experimental Data

In addition to the synthetic data sets, we evaluated all methods including FIRMA on two well studied cancer data sets. All exon array data used are from published work and are publicly available as stated in the corresponding articles.

The first data set is provided by Affymetrix [Gardina et al., 2006] and consists of 20 arrays, 10 colon cancer samples as well as their paired control. DS results were partly validated by RT-PCR. As a positive control (TP) we used all 18 probe set IDs indicated in the section 'differentially spliced between tissue types' and one additional probe set from the section 'previously reported splicing events in colon cancer' (see supplementary material [Gardina et al., 2006]) that was positively validated. The negative control (TN) was formed by the 10 probe set IDs in the section 'alternatively spliced but not differential between tissue types' (see supplementary material [Gardina et al., 2006]). Mapping to our data (we used only core exons and the human genome version 19) led to 12 TP and 8 TN probe sets corresponding to 10 (TP) and 8 (TN) genes respectively.

We also applied all methods on a lung cancer data set [Langer et al., 2010] consisting of 36 paired samples, 18 normal and 18 non small cell lung cancer (NSCLC) samples. The study provides validation data for 3 TN and 19 TP examples of DS.

Preprocessing and normalization of the data sets was performed as proposed in [Rodrigo-Domingo et al., 2013]. Background correction according to Irizarry [Irizarry et al., 2003b] was applied followed by quantile normalization. Summarization on exon and transcript level was done according to [Irizarry et al., 2003a]. To reduce the number of false positives due to a high number of statistical tests, 'absent' transcript clusters and probe sets were removed. Specifically, probe sets being absent in more than half of the samples of a group as well as transcript clusters with more than half of the probe sets absent were removed.

Evaluation

For the evaluation we determined, for each scenario, accuracy (ACC) or the area under the curve (AUC) in cases where the method yielded a continuous score. Furthermore, we quantified sensitivity and specificity for more fine-grained insights. Note that in the case of binary classification (DS event / no DS event) accuracy corresponds to the AUC.

Some of the methods produce p-values indicating the certainty of a DS event taking place, while PAC, KLAS, ARH and SI output a heuristic score. To achieve comparability and avoid cutoff problems, we also derived a p-value for all score-based methods using an exact Monte Carlo permutation test [Fay and Shaw, 2010]. Applied to the scores, a gene wise p-value is computed with a significance level of $\alpha = 0.05$. Nevertheless, we quantify performance on the basis of scores as well.

As stated, score based methods exhibit the difficulty of choosing a cutoff at which a result is believed to be relevant. There are best practices for some methods (SI is

mostly used with a cutoff of 1.5 [De La Grange et al., 2010] or 2 [Wang et al., 2009a]) or recommendations for others (ARH = 0.03 [Rasche and Herwig, 2010]) yet no appropriate value is known for PAC and KLAS. We therefore add a second evaluation for the score-based methods only based on AUC.

3.4 Results

Firstly, we report on the results for simulated data. The examined parameters (see section 3.3) were evaluated by p-value for all methods as well as by score for the score based methods only. Analysis of variance was applied to determine the significance of parameter influence.

Subsequently, the results on the colon and lung cancer data sets were reported with a focus on the RT-PCR validated results. As in the case of simulated data, accuracy, sensitivity and specificity was used to evaluate performance.

3.4.1 Synthetic Data

An overview on the accuracy for all scenarios, i.e. 16 synthetic data sets, is shown in Figure 3.2 using hierarchical clustering (euclidean distance, complete linkage) of methods as well as scenarios. The method performing best for each scenario is indicated by an asterisk, multiple maxima per column are possible. The most striking observation is the clear superiority of MADS', which performed equally well independent of data-imposed challenges. While most of the methods achieved good results in the 'easy' cases of equal group size and consistent splicing events, accuracy droppes quickly when sample sizes in groups diverged, less samples per condition were spliced, or expression intensity decreased. MADS' is closely followed by ARH, SI, SplicingCompass and KLAS, which showed similar behavior. The third-best cluster of methods consists of ANOSVA and MIDAS. The two performed well in the easy scenarios of sufficient sample numbers and 100% AS events in one group. As circumstances got more challenging, a rapid decay in accuracy could be observed.

Results per Method **MADS'.** This algorithm showed a unique performance not only concerning efficiency but also in the sensitivity to parameter influences (see Figure 3.3). The most obvious interference was incurred by expression level. While in the high expression range almost no FPs were observed, FP rate increased significantly in the scenarios with low expression. A second observation correlating with the expression level was the dependence on the number of exons contained in a gene. In low expression ranges MADS' performed consistently better in scenarios with a high number of exons per gene, while in high-expression scenarios it performed better with a low number of exons per gene.

ARH, SI, SplicingCompass and KLAS. The four methods behaved similarly in terms of classifying the genes actually spliced differentially (Figure 3.3). All showed a

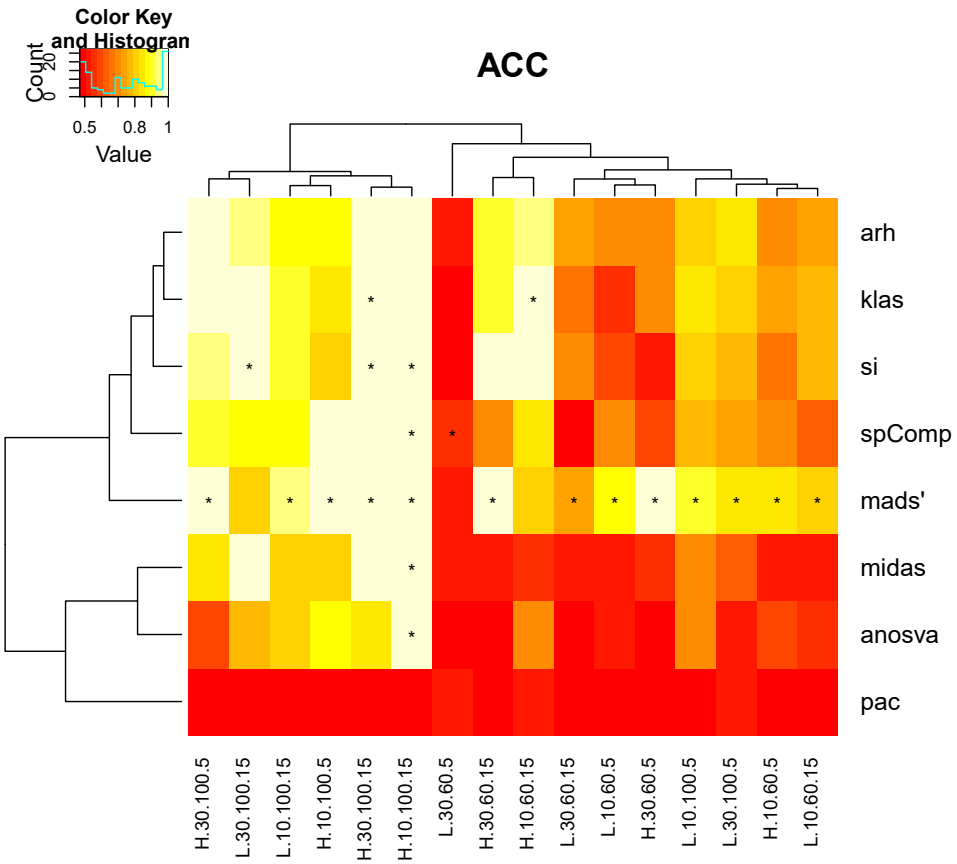
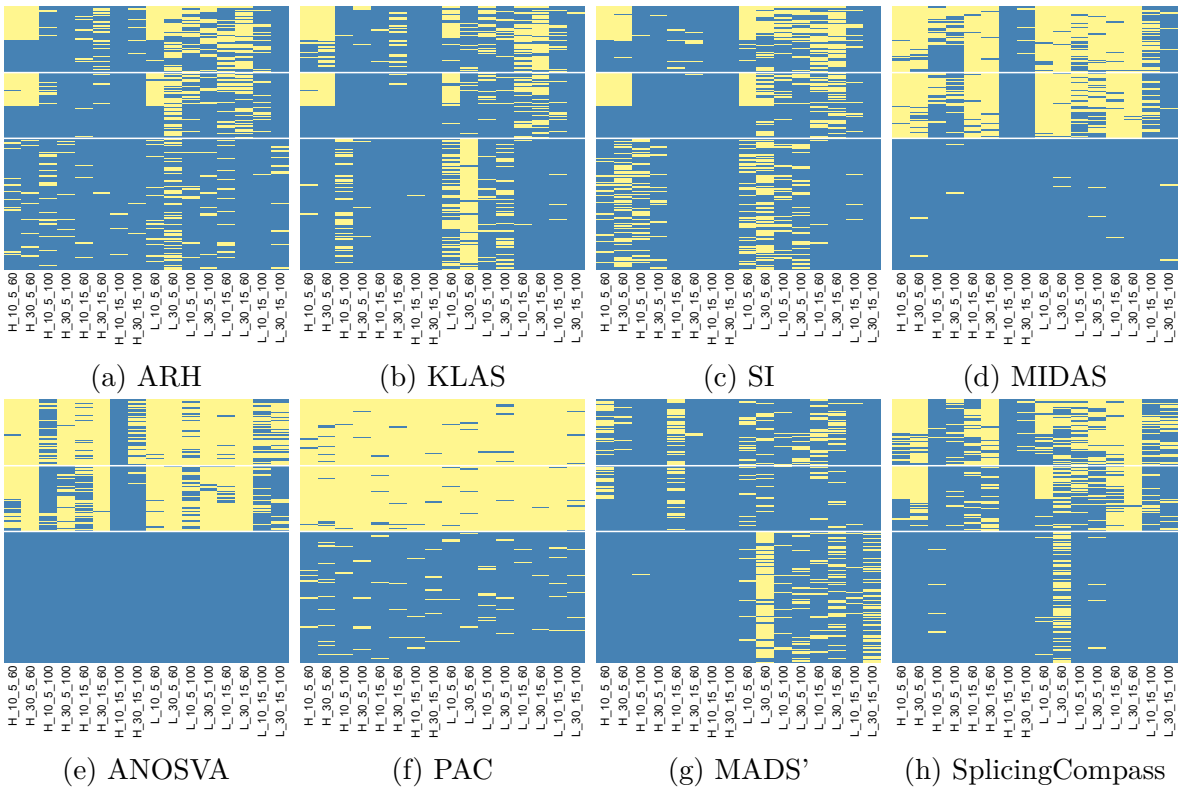


Figure 3.2: **P-value based accuracy** for all scenarios. Asterisks indicate highest values per scenario, multiple maxima are possible. Column names encode scenarios in the order expression.exons.percent.samples, thus H.10.100.5 describes the scenario with high expression, 10 exons per gene, 100 percent spliced samples in the respective group and 5 versus 15 samples per group. A light color indicates high, a dark (red) color indicates lower accuracy.

clear performance advantage in the case of high expression and they also shared the outliers: in the scenarios with 60% DS events and low sample size, genes containing the DS event in the small sample group were mostly not classified correctly. All other methods performed homogeneously bad or well irrespective of the fact that the DS event was not contained in the majority class. While ARH displayed a rather homogeneous response for the control genes, SI was strongly impacted by the number of samples per group. SplicingCompass displayed the lowest number of FPs in this group, as the consideration of all pairwise angles requires relatively strong effect sizes. Systematic influences observable by Figure 3.3 were exon number and percentage of samples displaying differential splicing.



(a) ARH (b) KLAS (c) SI (d) MIDAS

(e) ANOSVA (f) PAC (g) MADS' (h) SplicingCompass

Figure 3.3: **Classification by method** is displayed for all genes simulated. The upper half of the genes contain differential splicing events while the lower half serves as a control. Correctly classified genes are indicated in blue while incorrect predictions are highlighted in yellow. Genes in the first half of the DS set contain one while the second half contains two differentially spliced exons. In the scenarios displaying a different sample number per group and less than 100 percent differentially spliced samples the group containing the DS event is switched for half of the genes. Column names encode scenarios in the order expression_exons_samples_percent, thus H_10_5_100 describes the scenario with high expression, 10 exons per gene, 5 versus 15 samples per group and 100 percent spliced samples in the respective group.

ANOSVA, MIDAS and PAC. These methods formed the third method-cluster showing results very similar to each other throughout all scenarios. While ANOSVA and MIDAS were highly specific, ANOSVA educed not a single FP at the cost of a slightly lower sensitivity compared to MIDAS (see Figure 3.3 and Figure 3.5). The most obvious difference between the two was the difficulty of ANOSVA to deal with a high number of exons. MIDAS, on the other hand, performed independently of this parameter. As expected from a statistical method, the parameter impacting the performance most was the percentage of samples displaying the DS event in one group. Both methods failed to detect the DS event in most of the TP cases. Thus, if avoiding false positives is of high importance, MIDAS and even more, ANOSVA, are a suitable choice. PAC failed

to detect most of the positive events and also led to some FPs independently of the underlying scenarios.

Sensitivity and Specificity Depending on the aim of a potential study it can be important to choose a method explicitly focusing on high sensitivity or high specificity. While the first assures the correct detection of a sample having a certain property, the latter describes the ability to not detect samples not having this property, i.e. a certain disease. High sensitivity is required in all areas of diagnostics; when it comes to biomarker detection, a high specificity might be of higher interest. As biomarkers are usually used for screening of large populations for preventive reasons, a high number of false positives could lead to an increased workload of testing or unnecessary treatment [Hartwell et al., 2006].

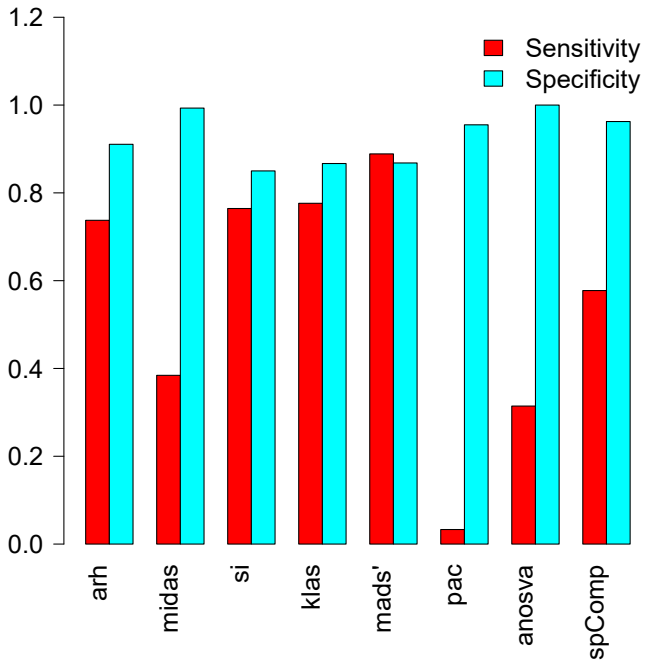
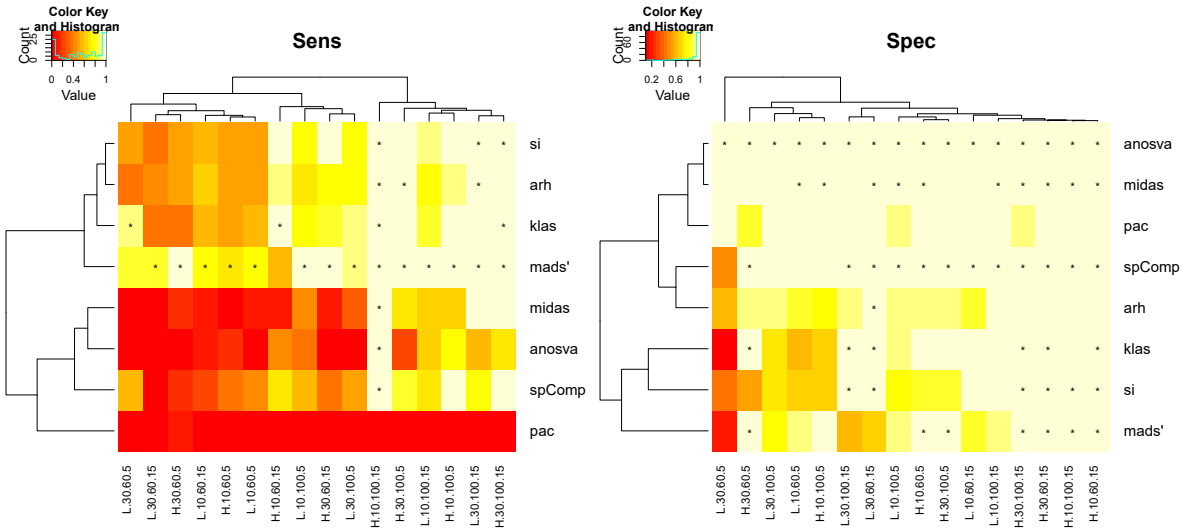


Figure 3.4: **Sensitivity and Specificity** for all methods applied on artificial data and averaged over scenarios.

Sensitivity and specificity for all scenarios are displayed in Figure 3.4 and Figure 3.5. High specificity values for ANOSVA, PAC and MIDAS came at the cost of sensitivity. While SI and KLAS presented very similar values - with KLAS showing a slightly better result - ARH was more focused on specificity.

SplicingCompass shows very high specificity yet lower sensitivity. Figure 3.5 gives a scenario-wide overview on specificity and sensitivity. Sensitivity was clearly dominated by MADS', followed by KLAS, exposing its strength in this category in comparison to its cluster mates.



(a) Sensitivity (b) Specificity

Figure 3.5: **Sensitivity and Specificity** for all scenarios (p-value based evaluation). Asterisks indicate the highest value(s) per scenario. Column names encode scenarios in the order expression.exons.percent.samples, thus H.10.100.5 describes the scenario with high expression, 10 exons per gene, 100 percent spliced samples in the respective group and 5 versus 15 samples per group.

Significance of Parameter Influence To access parameter influence in a systematic way, we fitted a linear model to the computed accuracy with a subsequent analysis of variance. The computed p-values indicate whether single parameters or combinations of two parameters have a significant influence on accuracy. Results are shown in Table 3.2 and Figure 3.6.

The greatest influence on the performance of all methods was the percentage of samples displaying a differential splicing event in one group. DS events contained in 100 % of the samples in one of two conditions led to consistently better results than in the 60% case. The highest impact was observed in MIDAS, KLAS, SplicingCompass and ANOSVA, all of them except for KLAS statistical approaches taking variance across samples into account.

A huge influence on performance was also notable for expression intensity. All methods except PAC showed a significant dependency on this parameter. The higher the expression, the easier was the distinction of DS events from background. Methods virtually not impacted by number of samples per group were MADS' and PAC. On the other hand, ANOSVA, KLAS, ARH, SI and MIDAS were significantly influenced by this parameter.

The number of exons per gene showed no impact on the performance of most of the methods. Two exceptions were ANOSVA and SplicingCompass, being consistently

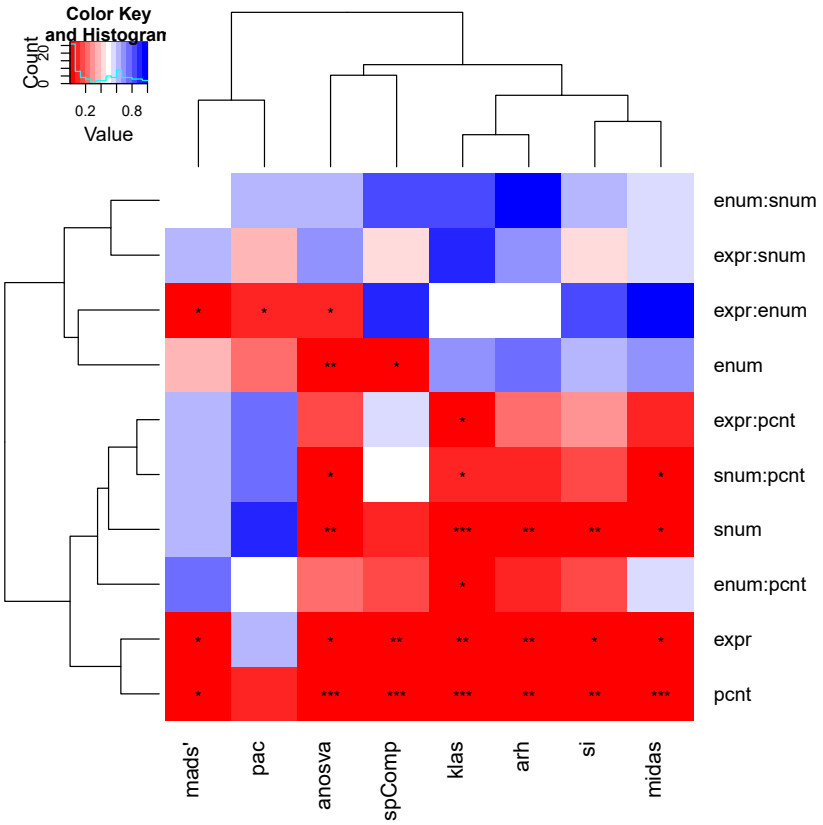


Figure 3.6: **Heatmap of ANOVA-based p-values.** Asterisks indicate significant values ($*** < 0.001$, $** < 0.01$, $* < 0.1$), red represents low, blue represents high p-values. Analysis of variance reveals the influence of the parameters as well as the influence of parameter combinations on the performance (i.e. accuracy). *enum*=number of exons, *snum*=number of samples, *pcnt*=percentage of differentially spliced samples in one condition, *expr*=expression intensity

superior in the case of lower exon numbers per gene across expression intensity variation. MADS' on the other hand showed a contrasting behavior in the high expression intensity (EI) (better results for low exon numbers) compared to low EI (better results for high exon numbers).

Influence of parameter combinations. Two out of the six combinations, i.e., *enum* : *snum* , and *expr* : *snum* showed no impact on the performance of the methods (see Figure 3.6). The joint effect of *pcnt* with *expr* and *enum* showed a slight influence on some methods, but only the one on KLAS was significant (Figure 3.6). A higher impact could be observed in the combination of *pcnt* and *snum*, which significantly influenced ANOSVA, KLAS and MIDAS and had a notable effect on ARH and SI as well. The collective impact of *expr* and *enum* had a significant influence on MADS',

	ARH	SI	KLA	MAD	MID	ANO	PAC	SCO
snum	+	+	+	-	+	+	-	-
enum	-	-	-	-	-	+	-	+
expr	+	+	+	+	+	+	-	+
pcnt	+	+	+	+	+	+	-	+

Table 3.2: **Analysis of variance reveals the significant influence of parameters on accuracy.** '+' indicates a significant influence of the parameter on accuracy, '-' means no significant influence.

PAC and ANOSVA while no effect was observed for KLAS, ARH, SI, SplicingCompass and MIDAS ($p - value > 0.05$).

3.4.2 Score Based Evaluation

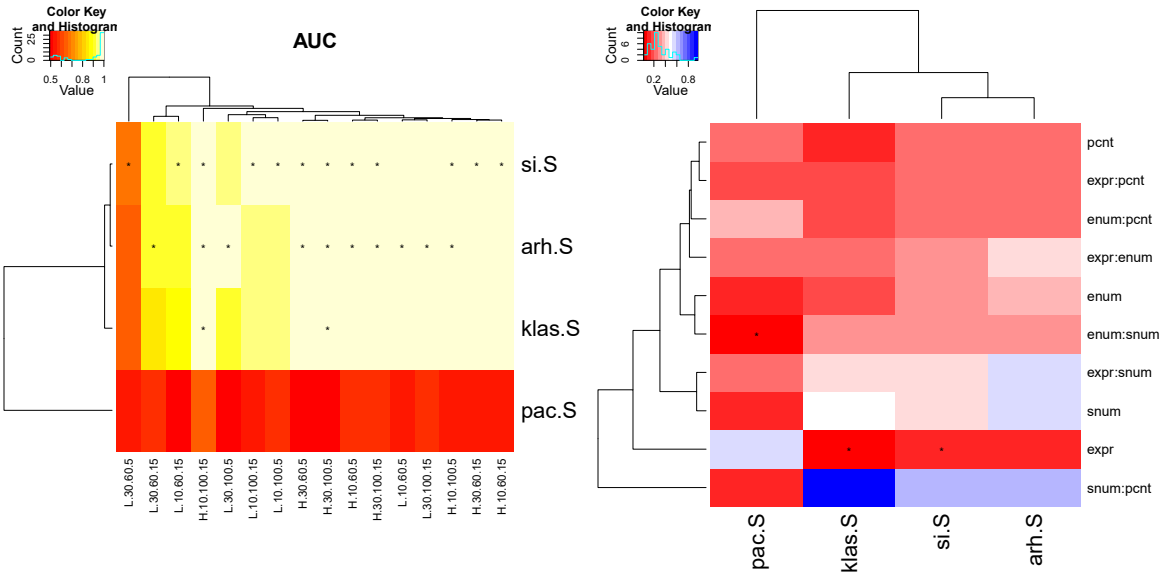
Besides the p-value based approach we were interested in comparing the ranking ability without having to decide on a cutoff. Thus, for each method we used the scores for computing the AUC. This led to a good performance of most of the methods (see Figure 3.7 left), emphasizing only marginal differences. SI and ARH showed a slight superiority in performance to KLAS, and were thus - for this scenarios - favorable over the latter when relied on scores only.

Dissecting the score based results by parameter (see Figure 3.7 right) revealed a dependency on the expression level for KLAS and SI as well as a significant impact of the combination of exon number and sample number on PAC. For high expression, most score based methods performed rather good, independently of the percentage of differentially spliced genes per group, the number of samples per group or the number of exons per gene. When it came to a lower expression level, performance in the 'harder' scenario decreased. This phenomenon was most obvious for the percentage of differentially spliced genes per group.

3.4.3 Experimental Data

We applied all nine methods - including FIRMA - to two partly RT-PCR validated data sets, one consisting of colon cancer and the other of lung cancer samples. First, we investigated the overall predictions of every method to assess the number of prognosticated differential DS events. Second, we compared the predictions based on TPs and TNs confirmed by RT-PCR. The p-value cutoff is set to 0.05.

Colon Cancer ARH, KLAS and ANOSVA produced approximately the same gene number (about 2000) while slightly differing in the gene set. SI and FIRMA proposed about 1000 differentially spliced genes while PAC, MIDAS and Splicing Compass showed the most conservative result (less than 500 genes). MADS' predicted the highest number



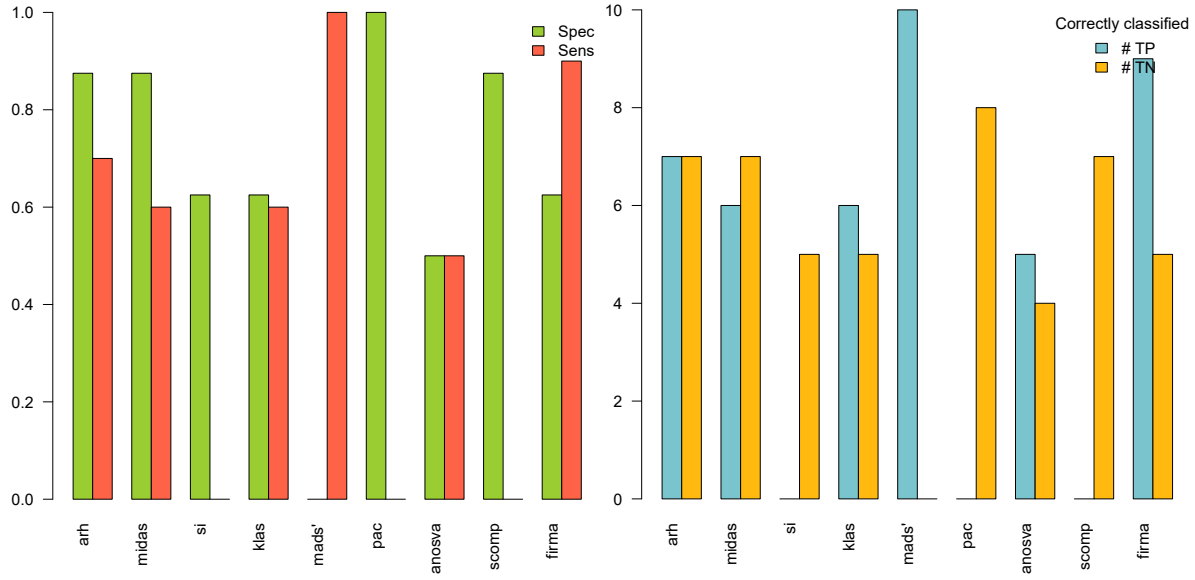
(a) AUC (b) P-values

Figure 3.7: **Score based AUC for all scenarios (left)).** Asterisks indicate the highest values per scenario. Column names encode scenarios in the order expression.exons.percent.samples, thus H.10.100.5 describes the scenario with high expression, 10 exons per gene, 100 percent spliced samples in the respective group and 5 versus 15 samples per group. **Heatmap of ANOVA-based p-values (right).** Asterisks indicate significant values ($*** < 0.001$, $** < 0.01$, $* < 0.1$). Analysis of variance reveals the influence of the parameters as well as the influence of parameter combinations on the performance. (enum=number of exons, snum=number of samples, pcnt=percentage of differentially spliced samples in one condition, expr=expression intensity)

of DS events (> 13000) (Figure 3.9) and thus clearly overrates DS in our adaption of the method.

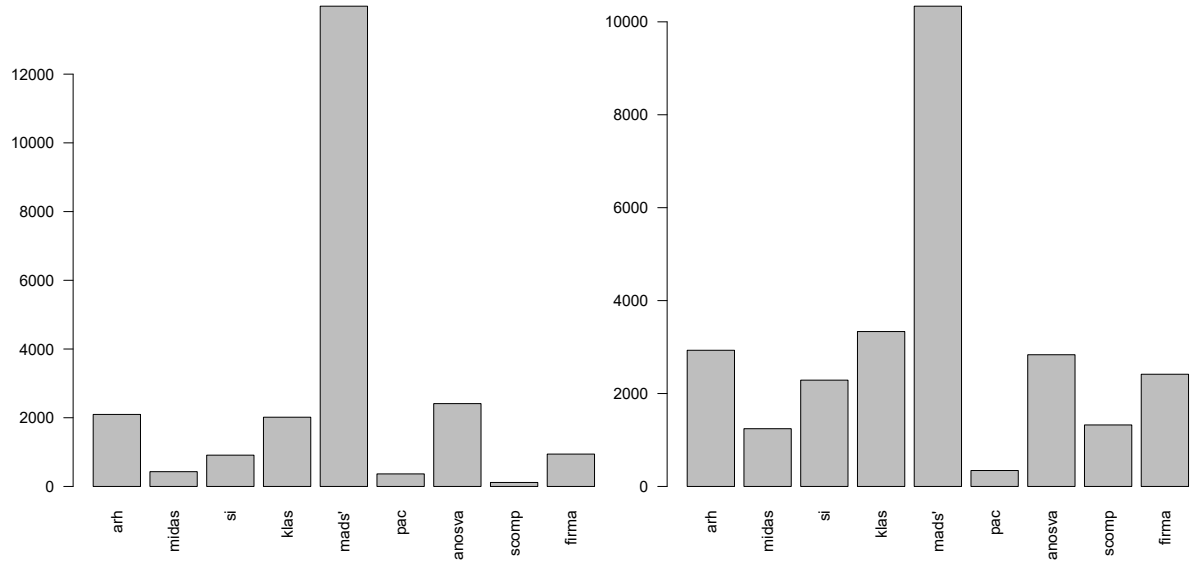
Thus, MADS' was an outlier in the number of predicted DS events, claiming the sought event in over 70% of the genes. When considering only the validated results ARH and FIRMA appeared as the most accurate methods (see Figure 3.11) closely followed by MIDAS. KLAS and ANOSVA displayed relatively good results whereas the remaining three methods showed either a high sensitivity at the cost of specificity (MADS') or a high specificity with a sacrifice of sensitivity (SplicingCompass, PAC), see Figure 3.8.

Lung Cancer ARH, KLAS, FIRMA, and ANOSVA predicted about 3000 DS events with considerable overlap in the gene set. SI nominated about 2000, MIDAS and Splicing Compass 1000 and PAC showed the most conservative result with less than 200 genes. Again, MADS' predicted the highest number of DS events (> 10000) (Figure 3.9). As the data set provided such a high verification rate, number of TN examples was very



(a) Sensitivity and Specificity (b) RT-PCR validated DS events

Figure 3.8: **Sensitivity and specificity for RT-PCR validated DS events** in the colon cancer data set (left) and the **number of RT-PCR validated DS events** for every method (right).

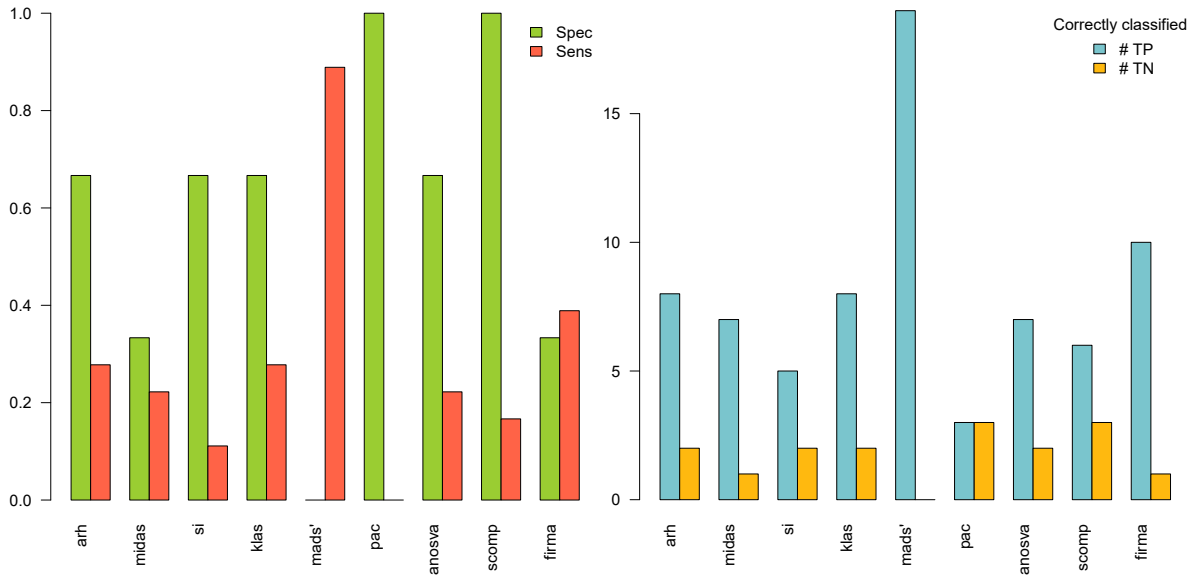


(a) Colon Cancer (b) Lung Cancer

Figure 3.9: **Number of genes being predicted as differentially spliced per method** for the colon cancer data set (left) and the lung cancer data set (right).

low (we used non-verified events as TN). Under such circumstances accuracy is not a

good measure for performance, and we thus focused on sensitivity and specificity instead (Figure 3.10).



(a) Sensitivity and Specificity (b) Number of TP and TN

Figure 3.10: **Sensitivity and specificity for RT-PCR validated DS events** in the lung cancer data set (left) and the **number of RT-PCR validated DS events** for every method (right).

SplicingCompass, ANOSVA, KLAS, FIRMA, and ARH were the methods performing best. According to accuracy, FIRMA, KLAS and ARH achieved the highest values when disregarding MADS' due to its high prediction rate. Similarly, sensitivity was also dominated by FIRMA, KLAS and ARH while considering only methods with non-zero specificity values. When focusing on specificity, SplicingCompass was the clear winner followed by ANOSVA, KLAS, ARH and SI, all ranging on the second place.

3.5 Discussion

Though a variety of methods for the detection of DS based on exon array data has been developed over time, no broad evaluation concerning their advantages and drawbacks in regard to (combined) influences of properties such as the number of samples, expression intensity or exon number has been performed yet. In this work we evaluated the impact of an extensive set of parameter combinations on the performance of eight methods. Additionally, we assessed all methods and a ninth one with respect to validated experimental data. In contrast to related work which focused on the comparison based on experimental data [Rasche and Herwig, 2010] and thus on fixed scenarios, we also exploited simulated data sets to study the (combined) influence of various properties of differentially spliced genes and their measurements in exon arrays.

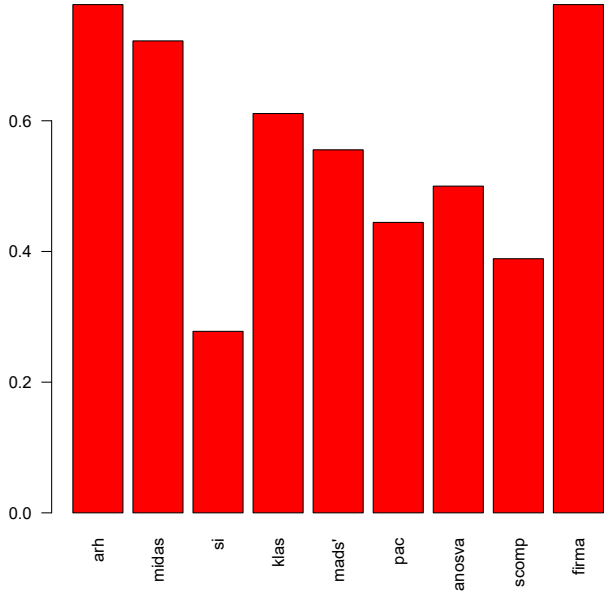


Figure 3.11: **Accuracy** computed on the RT-PCR validated results for the colon cancer data set.

A rank comparison of accuracy-based results is shown in Table 3.3, putting results on synthetic and on real data sets side-by-side together with the ranking reported in [Rasche and Herwig, 2010]. Here, we present the outcome of synthetic, colon cancer and lung data. Concerning the accuracy based results, some methods ranked consistently low (ANOSVA, SI, SplicingCompass and PAC), others consistently high (ARH, FIRMA and KLAS) while MIDAS included a positive outlier. Note that we did not include MADS' in this rank comparison, as results on experimental data suggest an overrating of DS events.

Nevertheless, results on experimental data should be handled with care due to the unbalanced nature and small size of the evaluation data in these data sets. Recall that accuracy is highly susceptible to a diverging number of positive and negative examples. Especially in the case of MADS', which predicted a high number of differentially spliced genes, combined with an disproportionate high number of positive examples in the lung cancer data set this is an issue.

3.5.1 Algorithmic Performance

Clearly, the performance of different algorithms was influenced differently by the various parameters and the different data properties, such as effect size and variance. To shed more light into the cause of these differences, we here sought to explain differences in the method's performance in terms of their underlying mathematical formulation of the problem.

	Accuracy				AUC
	Synthetic	Colon	Lung	Median	Rasche
ANOSVA	6	4	4	4	4
ARH	1	1	3	1	1
FIRMA	n.a.	1	1	1	5
KLAS	2	3	3	3	n.a.
MIDAS	5	2	6	5	6
Splicing Index	3	7	7	7	2
SplicingCompass	4	6	5	5	n.a.
PAC	7	5	8	7	3

Table 3.3: **Result summary and comparison.** Per dataset D and method M we show the rank that M achieves on D, when all methods are sorted by accuracy, i.e., the number of truly recognized splicing events. For comparison, we also add ranks from Rasche et al. [Rasche and Herwig, 2010], which used a different data set and ranked by AUC.

Exon Number and Differentially Spliced Exons per Gene Two methods - ANOSVA and SplicingCompass - were significantly affected by the number of exons per gene, i.e. they display a better performance in the low exon number scenario. This is remarkable, as a major concern of most other algorithms is a rising number of FPs with increasing exon number due to parallel tests. In the case of ANOSVA, the reason is that, the higher the number of exons, the more improbable it becomes to obtain significant predictions for TPs as the number of differentially spliced exons remains constant. This is underpinned by the observation that predictions were better in the second half (genes 50 to 100) of TPs, where two instead of one exon is modeled as differentially spliced. The same reason applies to SplicingCompass, a statistics-based method, which accesses the difference between exon angles within and between groups. The higher the number of exons - while the number of DS events is constant - the lower is the ratio of angles representing a DS event. This impedes the detection of differences between groups.

Interestingly, MIDAS, also a statistical method, was not affected by these parameters. Unlike SplicingCompass and ANOSVA, MIDAS directly takes into account the gene expression normalized exon expression, i.e. effect size, and applies a separate test for every exon. The number of exons per gene is thus not as important. In contrast, SplicingCompass and ANOSVA operate on a gene-based level.

Sample Number and Variance For any method based on statistical tests, one expects that a higher number of samples improves performance as it increases test power. As expected, this behavior was observed for ANOSVA and MIDAS, both inherently statistical methods. However, the same (positive) effect also could be observed for SI, ARH and KLAS, which do not perform tests. The explanation is that all these three methods

use permutation tests, which become more stable with increasing numbers of samples. The effect was the strongest for SI with those genes which were not differentially spliced (see Figure 3.3).

Expression Level and Effect Size All methods were significantly affected by the expression level: The lower, the worse were the results. This is to be expected, as low expressions means a less clear separation between signal and noise. As expression decreases, also the variance decreases, which in turn makes it more probable to confuse spurious 'effects' as splicing events.

MADS', for instance, showed this behavior for the non differentially spliced genes, by producing a high number of FPs in the low expression scenarios which is not visible for similar methods, like for instance MIDAS. While MIDAS computes an exon level SI and subsequently applies statistical testing, MADS' produces a gene wise aggregate as final p-value. The approach of MADS' is thus more sensitive and yields performance improvements but can, on the other hand, also be too sensitive for other scenarios (e.g., see Figure 3.3).

The rather simple splicing index performed well in most of the scenarios, although this method does not consider variance and does not perform any kind of deviation correction. However, this is due to the structure of the generated data, while various influences alter the challenges imposed by the data, the one affecting SI most - a small number of rather drastic outliers - was not contained in the scenarios. Thus the focus on effect size led to remarkable results.

Percent of Spliced Samples The greatest impact due to this parameter is observed for statistical methods, i.e. ANOSVA, MIDAS and SplicingCompass. As they are by design susceptible to variance, fluctuations like in the case of decreased sample ratio with DS events per group (i.e. a lower percentage of differentially spliced samples) lowers performance as increased variance prevents effects from being significant.

Effect Size, Variance and Gene Level Correction As already mentioned in the previous paragraph, statistical methods in general are rather conservative in predicting DS events. One root of this behavior is their test-basis, but other effects come on top. MIDAS uses gene-normalized expression values instead of exon expression values and thus requires a fairly great effect as the normalization is rather drastic. ANOSVA applies an ANOVA on a so-called interaction term derived from a fitted linear model which further smooths away differences. Other methods are less strict in these regards. For instance, ARH uses the median exon ratio between groups for correcting for the underlying gene expression. Compared to MIDAS, which directly uses exon to gene ratio, the approach of ARH often results in a less pronounced correction which better preserves effect strength. Splicing Compass accesses the difference between exon angles within and between groups. It does not perform any explicit gene level correction, but implicitly all pairwise angels are considered, resulting in an indirect and rather weak form of normalization. Again, this helps this method to increase its sensitivity.

The ambivalence of MADS'. Combining the results of simulated and experimental data completes the picture of MADS'. While leading performance for simulated data, MADS' seemed to overrate DS events in the experimental settings. The excellent performance in the artificial scenario reflects the strong sensitivity of the method: relatively 'hard' scenarios are still positively identified, settings in which other methods clearly voted against an DS event. According to our experiments MADS' can not be recommended for the pure prediction of DS events, but we consider it highly suitable for ranking DS candidates because genes with a very low MADS' p-value very likely show differential splicing.

Which Method for which Data? Depending on the research question and the experimental data, different methods pose an appropriate choice. As practically all methods showed a significant dependency on the expression level and the amount of differentially spliced samples per class the two parameters are of no help for method selection. If sample number is low and / or imbalanced, SplicingCompass is the most reasonable choice according to our evaluation. Independence on the number of exons is best achieved by ARH, while KLAS, SI and MIDAS pose similarly good choices. High specificity throughout the data sets was provided by ARH, SplicingCompass and MIDAS. When it comes to the most sensitive methods FIRMA, ARH and KLAS fulfill the task best. As validation of results is expensive and time-intensive most studies are interested in high sensitivity and specificity as well as in robustness of the method. According to our evaluation, ARH meets these requirements best.

3.5.2 Comparison to Related Work

A comparison of MIDAS, FIRMA, SPLICE [Hu et al., 2001], ARH, PAC, SI, ANOSVA, MADS, and correlation [Shah and Pallas, 2009] has been performed previously [Rasche and Herwig, 2010]. However, the evaluation of Rasche et al. used only a single scenario by benchmarking on different tissue data, while our main interest lies in the susceptibility of the methods to different data properties. Furthermore, [Rasche and Herwig, 2010] focused on ranking performance and evaluated based on AUC instead of accuracy, sensitivity and specificity. Using AUC avoids the problem of choosing a cutoff, but precisely the proper selection of a cutoff decides on the usefulness of a method in reality. Due to such differences, a comparison of our results with those from [Rasche and Herwig, 2010] should be interpreted carefully as the two measures quantify a different matter.

The most striking difference is the good performance of PAC. PAC strongly depends on the gene estimate and the exon estimate used. Furthermore, we compute p-values from PAC scores, which were much more susceptible to noise than for example the SI and therefore had difficulties leading to significant results. Further comparative work was done by Laajala et al. [Laajala et al., 2009]. Though focusing on preprocessing, they implicitly compared FIRMA, SI and MIDAS, indicating that MIDAS develops its strength with growing number of differentially spliced exons.

3.6 Conclusion

Over time a variety of methods for the detection of DS has been published, each of them with different characteristics regarding sensitivity, specificity, interference to certain data settings and robustness over multiple data sets. In this work, we discerned the various methodological aspects for the first time. Using synthetic data complemented by validated experimental data enabled us to obtain a thorough overview on the performance of the individual methods, including advantages and drawbacks, which is essential for avoiding a misinterpretation of the respective outcomes.

While some methods, such as ARH, perform consistently well over all data sets and scenarios, other methods show heterogeneous prediction quality on the different data sets. The adequate method has to be chosen carefully and with a defined study aim in mind to prevent, for instance, a high amount of FPs when validation is laborious.

To avoid an unfeasible flood of data scenarios we restricted our simulations to cases, where one and two exons are differentially spliced per gene. Naturally, this does not represent the spectrum of actually occurring DS events. Thus, an important question to address in future work is the susceptibility of the methods to the number of differentially spliced exons per gene. Further insights into the robustness of a method can be obtained by varying the noise level during data generation.

Nevertheless, we provided a systematic and elaborate overwiev on a variety of properties for DS detection methods, enabling researchers a targeted selection to address the fundamental topic of differential splicing.

4 Splicing Factor Network

Splicing is a crucial mechanism for establishing eukaryotic protein diversity which is regulated by splicing factors (see Section 2.1.4). Malfunctioning of these regulatory proteins may yield aberrant protein isoforms involved in the onset of various diseases such as cancer. While transcriptomic data enable the identification of differentially spliced exons (see Chapter 3), the cause of aberrant isoforms often remains unclear, as transcriptional changes in splicing factors are usually minor and thus difficult to detect. In some cases there are no changes at the transcriptional level at all, as modifications, such as phosphorylation, might interfere with the efficiency of a splicing factor.

Here, we aim at identifying splicing factors most probably responsible for changes in splicing observed between a lymphoma subtype and a control group. To this end, we developed a network-based approach, ranking known splicing factors according to the evidence of them causing the observed DS events. We apply our approach to exon expression data derived from 113 patients in six lymphoma subtypes and a non-malignant control group (see Section 2.3.2).

4.1 Introduction

Splicing is a complex and still not completely understood process involving various players in numerous combinatorial settings [Cáceres and Kornblihtt, 2002]. A detailed description of the regulation of alternative splicing is given in Chapter 2.1.4. Two fundamental types of components involved in this process are cis- and trans-acting factors. Cis-acting factors, also referred to as splicing regulatory elements (SRE) are part of the genetic code, and located in promoter regions, exons or introns.

SREs interact with trans-acting factors, also known as splicing regulatory proteins or splicing factors (SFs). Published estimations on the number of SFs in human vary, depending on the exact definition, between 70 and 250 [Korneta et al., 2012, Wang and Burge, 2008, Giulietti et al., 2012, Agafonov et al., 2011]. They recognize certain cis-acting elements due to their binding site and regulate and control splicing [Zhou et al., 2002] through intercommunication and a subtle and elaborate equilibrium of their abundances [Dredge et al., 2005, Gunderson et al., 1997]. Perturbations in this fragile system may lead to a cascade of changes some of which are causal for various diseases. Amongst the known malfunctions, up to 30 % are involved in the onset of cancer [Xi et al., 2008, Ghigna et al., 2008]. Elucidation of (1) the precise aberrant splicing events as well as (2) the regulatory malfunctioning causal for the change in exon inclusion in a given sample is therefore crucial for understanding effects and origin of cancer.

The underlying causes for these perturbations can be manyfold. Besides changes in expression, other effects such as auto-regulation of SFs [Dredge et al., 2005, Gunderson et al., 1997], epigenetic modifications [Dardenne et al., 2012, Luco et al., 2011, Watson et al., 2013, Zhou et al., 2012] as well as post-translational modification mechanisms such as phosphorylation [Wang et al., 1999, Liu et al., 2013, Stamm, 2008] and ubiquitination [Moulton et al., 2014, Bellare et al., 2006] contribute to the regulation of alternative splicing.

Detecting alterations on all these levels is a time-intensive and financially demanding task. Here, we seek to identify SFs involved in the observed splicing differences between two conditions irrespective of the causal mechanism. Our main idea is to indirectly analyze such effects using only transcriptome data. This is done by associating observed splicing changes on the transcriptome level with all known SFs as potential candidates responsible for the changes observed. To this end, we construct a network for each of the two conditions, both containing all exons differentially spliced between the conditions as well as all potentially causal genes, i.e. splicing factors. Thus, both networks initially contain the same nodes. Edges are labeled with condition-specific expression correlation and removed when under a certain cutoff to assure high confidence. SFs in these two networks are then ranked according to their centrality and SFs differeing most between the networks are candidate SFs for implications in condition-specific splicing [Lichtblau et al., 2016].

By combining the potentially causal candidates, i.e. SFs, with the known changes, the differentially spliced exons, we expect to detect not only differentially expressed SFs - whose identification can be considered a proof-of-concept - but also genes without changes in expression, yet with large network effects, that might help to link alterations on other genomic or biological levels, such as SNPs, phosphorylation or epigenetic changes, enabling a targeted investigation.

Network approaches have already been widely and successfully used in the biomedical application field [Aittokallio and Schwikowski, 2006] including splicing regulatory factors and the exons they control. Dai et al. [Dai et al., 2011] investigated splicing modules, i.e. exons being controlled by the same splicing factor. Chen et al. [Chen and Zheng, 2009] seeked to reveal associations between exons and their regulators and targets. A global approach is pursued by Qu et al. [Qu et al., 2010] predicting genome wide splicing regulatory networks (SRN) from public gene and exon array data. While Qu et al. reconstruct a healthy, genome-wide SRN we aim at specifically identifying the SFs related to splicing changes in lymphoma by using a healthy and a disease network.

We apply our approach to an extensive lymphoma data set comprising 113 samples, where expression is measured using exon arrays. For more details on the data set see Chapter 2.3. The samples originate from six different lymphoma subtypes as well as a tonsil control group. We therefore study DS in six lymphoma subtypes by comparing each of them to a non-malignant control group (see Table 4.1). For each of these comparisons, we construct two networks, one control (tonsil) and one disease (lymphoma subtype) network. The application of our method leads to candidate SFs which we extensively discuss and classify according to current knowledge.

4.2 Methods

In the following we describe the aquisition of all network components as well as the network analyses performed on the data.

4.2.1 Differential Expression Analysis

We analysed transcription data for 6 lymphoma subtypes and one control group (tonsils) measured with Affymetrix Exon Arrays (GeneChip Human Exon ST Array). For the number of samples contained in each subgroup as well as an introduction on the data see Table 4.1 and Chapter 2.3. Preprocessing of the data was accomplished according to [Rodrigo-Domingo et al., 2013]. Differential expression as well as differential splicing was determined as follows. Gene level analysis for all lymphoma subclasses compared to control (tonsil) was scored using limma [Smyth, 2005], whereby a fold change (FC) of 2 and a multiple testing (Benjamini Hochberg) corrected p-value [Benjamini and Hochberg, 1995] of 0.05 were used as cutoff. Based on our evaluation in Chapter 3, we decided to use the intersection of three methods, FIRMA [Purdom et al., 2008], ARH [Rasche and Herwig, 2010] and KLAS [Jentsch, 2011] to filter out less certain events.

The Splicing Factors we used for network construction are derived from SpliceAid-F [Giulietti et al., 2012], a public, hand-curated database for human splicing factors and their RNA binding sites. From the 71 SFs contained in the database, we could map 54 to our expression data. A full list is given in Appendix 6.1, Table 6.13.

4.2.2 Network-based Analysis

For every lymphoma subtype we constructed two networks, a control (Tonsil) and a disease network, both containing all splicing factors and all differentially spliced exons with respect to the considered subtype-control comparison (see Figure 4.1). Edge weights were determined by using Pearson correlation of all pairs. Thus, before applying a correlation cutoff for edges, the only difference between the two networks per lymphoma is the edge weights determined by the expression correlation.

In both networks, we ranked all splicing factors according to their betweenness centrality (BC) [Freeman, 1977] in the respective network. To determine which of them display the greatest changes in centrality, i.e. play a different role in the two networks, we compared the corresponding BC values to each other.

In this work, we are interested in a rather rank-independent difference as slight changes in the behavior of splicing factors can have a great impact on the splicing machinery. Neither absolute nor relative changes in centralities are taken into account quantitatively when assessed on rank difference solely. Thus, this is the least favorable option for the task. Relying on absolute changes on the other hand does not take into account the initial BC-rank of the entities. A top ranked candidate can have the same absolute difference as the last ranked candidate while their relation to each other is not displayed in this measure. Thus, we use the relative change of each entity, i.e. the difference

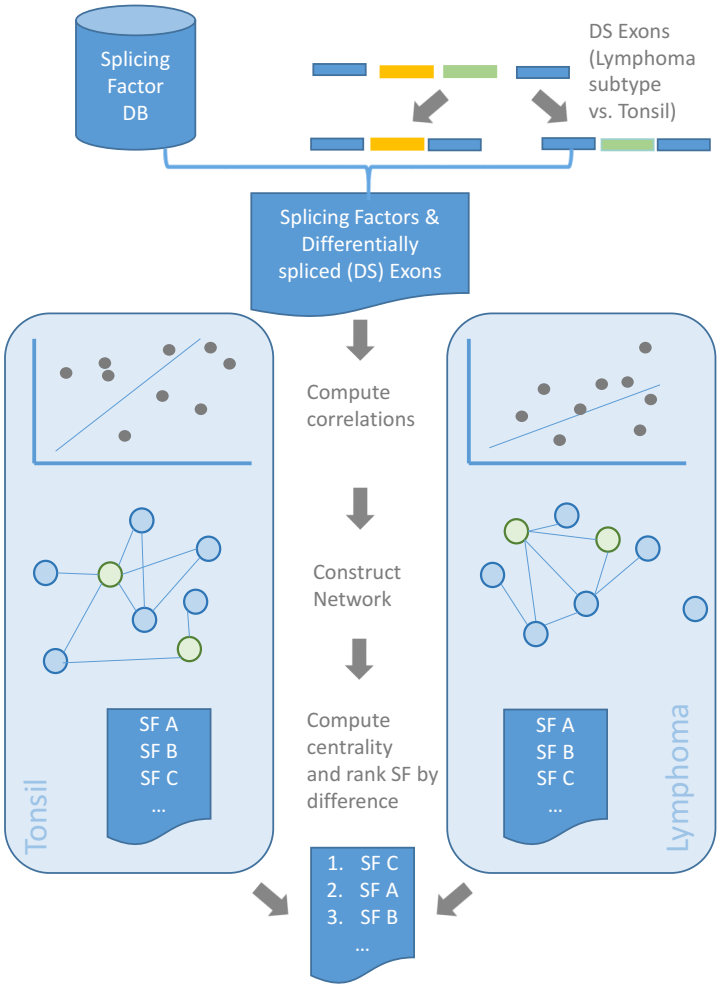


Figure 4.1: **Splicing factor network construction.** Schematic overview on the network construction and splicing factor ranking. For each comparison of lymphoma subtype versus control two networks are constructed both containing all SFs obtained from a public database and all differentially spliced exons between the two conditions. For each pair of network components (SF, exon) expression correlations are computed and used as edge weights. Subsequently, betweenness centrality for all SFs in both networks is computed. The difference in centrality between the networks is used for ranking of the SFs. The more different their centrality, the more likely we deem their involvement in aberrant splicing for the respective lymphoma subtype.

between logarithmic BC values. An entity doubling its centrality relative to the other network is now ranked equally independently of their rank in the networks.

Correlation networks change depending on the correlation cutoff used for their construction. To filter for the most stable results, we applied our method using eight different correlation cutoffs equally distributed between 0 and 1 (absolute values). For all cutoffs used, we determined the top ten differentially central SFs. We considered

those SFs as candidates per lymphoma that occurred in the top ten in at least half of the network comparisons.

4.3 Results

In the following section, we present the results for the application of our method. More precisely, we report the differential expression as well as the differential splicing identified for all subtype-control comparisons. Subsequently, we go into SF showing differential centrality (DC) and their expression behaviour. We discern functional implications of differentially altered entities, explore potential drivers behind candidate-SFs and investigate the possible role of SNPs for interesting SFs.

4.3.1 Expression Changes

The numbers of differentially expressed SFs and differentially spliced exons for each subtype are shown in Table 4.1. Whereas each subtype displays an impressive number of DS events, only very few splicing factors show differential expression. This observation supports the assumption that changes in SF expression are either too small to be detected by commonly used cutoffs or that alterations not based on expression level such as epigenetic modifications, mutations or post-translational modifications are involved [Luco et al., 2011, Brown et al., 2012, Zhou et al., 2012, Naro and Sette, 2013, Liu et al., 2013, Zhong et al., 2009, Xi et al., 2008]. Note that our network centrality based method captures these influences indirectly.

Samples		Differential Expression			Enrichment
subtype	number	DE SFs (pval & FC)	DE SFs (pval)	differentially spliced exons	fisher test pvalue
CLL	14	2	15	1395	0.43
DLBCL	40	1	14	1152	0.006
ALCL	10	1	7	801	0.0002
PTCL	6	1	3	337	0.05
MCL	12	0	2	496	0.35
FL	22	1	2	879	0.30
Tonsil	9	-	-	-	-

Table 4.1: **Sample and result overview. Samples** displays the number of samples for each lymphoma subtype as well as the tonsil control group. **Differential Expression** lists the number of differentially expressed SFs (compared to the control group), with and without fold chance cutoff, as well as differentially expressed exons per condition. **Enrichment** Fisher Test based p-value for enrichment of differentially expressed SF in top differentially central SFs.

4.3.2 Differentially Expressed SF Tend to be Differentially Central

Based on the idea that SFs with an altered expression between two conditions contribute to the splicing changes observed, we expect differentially expressed SFs to rank highly in the list of SFs ordered by differential centrality (but not necessarily vice versa). As only a small number of SFs meet the criteria of differential expression (FC > 2, p-value < 0.05), we also include SFs that do not meet the full criteria concerning FC, but display only p-values according to our restrictions (Table 4.1). For ALCL, p-value based differentially expressed SFs are shown in Figure 4.2.

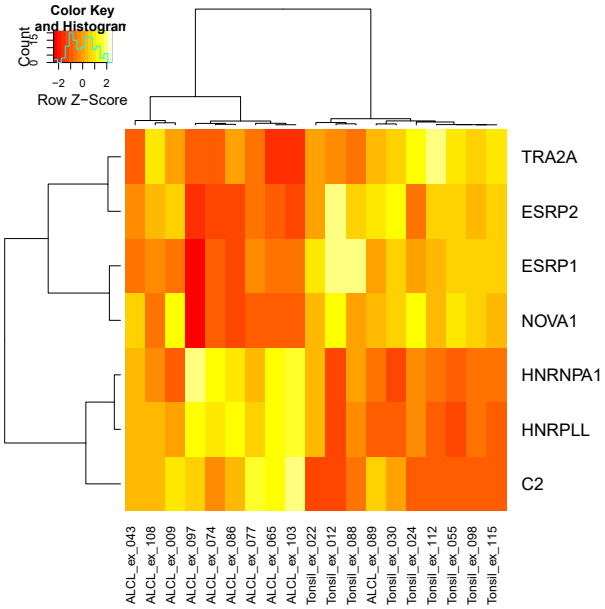


Figure 4.2: **Differentially expressed SFs in the comparison ALCL vs. Tonsil.** Hierarchical clustering of gene expression values corresponding to SFs with a significant p-value for the comparison of ALCL and Tonsil. Except for one outlier, ALCL_ex_089, a clear separation of the two conditions is visible.

Three out of six lymphoma subtypes showed a very low number of differentially expressed SFs, even when applying only the p-value as cutoff. Thus, we do not consider peripheral T-cell lymphoma (PTCL), mantle cell lymphoma (MCL) and follicular lymphoma (FL) as a good model to perform this type of validation. ALCL, chronic lymphocytic leukaemia (CLL) and diffuse large B-cell lymphoma (DLBCL) show seven to fifteen DE events and are thus better suited for this purpose. We conducted a fisher exact test to assess the significance of the top positioning of our candidates. Two out of these three conditions, ALCL ($p = 0.0002$) and DLBCL ($p = 0.006$), result in a highly significant enrichment of differentially expressed SFs amongst the top ranked differentially central SFs. CLL, the third condition shows no significant enrichment ($p = 0.43$).

4.3.3 Role and Function of Differentially Expressed SFs

SFs differentially expressed between conditions have high evidence of being causal for splicing changes observed. Thus, their functional characterization is of great interest for the elucidation of the underlying pathological mechanism. Remarkably, two splicing factors showed differential expression throughout most subtypes. These two, ESRP1 and ESRP2, are epithelial splicing regulatory proteins known to play a crucial role in splicing changes involved in epithelial-mesenchymal transition (EMT) [Warzecha et al., 2009a]. Several publications on the identification of potentially regulated splicing candidates exist [Dittmar et al., 2012, Warzecha et al., 2009b]. Amongst them is CD44 [Warzecha et al., 2009b], a protein well known in the context of splicing in lymphoma [Wallach-Dayan et al., 2001, Salles et al., 1993, Stauder et al., 1995, Yae et al., 2012]. Other well characterized targets are ENAH and CTNND1 [Warzecha et al., 2009b]. All three targets showed differential splicing in most of the conditions in our data as well.

Another SF showing differential expression (DE) in several conditions (ALCL, CLL, DLBCL) is NOVA1, a neural-specific RNA binding protein associated with paraneoplastic disorders [Buckanovich et al., 1996], and hepatocellular carcinoma [Zhang et al., 2014]. Interestingly, a role in some lymphoma cell lines is also reported [Relógio et al., 2005].

4.3.4 Differentially Spliced Genes and their Functional Implications

A remarkable number of commonly differentially spliced exons, especially in CLL, DLBCL and FL, can be observed in different lymphoma subtypes (see Figure 4.3).

Furthermore, a high amount of group specific splicing is detected for ALCL, DLBCL and CLL. For few subtypes, some SFs themselves showed differential splicing (see Table 4.2). This event occurred in three different conditions, while ESRP1/2 was present in most of the cases.

SF	CLL	DLBCL	FL
ESRP1	+	-	+
ESRP2	+	+	+
NOVA1	-	+	-
KHSRP	-	+	-

Table 4.2: **Splicing factors differentially spliced.** '+' indicates differential splicing, '-' indicates no differential splicing in the respective condition.

For an overview on the main functional involvements of the differentially spliced genes (Appendix 6.1, Tables 6.1 to 6.6) we applied a DAVID [Huang et al., 2008] analysis on GO Biological Process terms. Results are shown in Appendix 6.1 Tables 6.7 to 6.12. Most terms are attributed to differentiation and developmental mechanisms. Table 4.3 shows the intersection of the significantly enriched GO terms for the different subtypes. Interestingly, a high number of keratin subtypes rank among the top differentially spliced

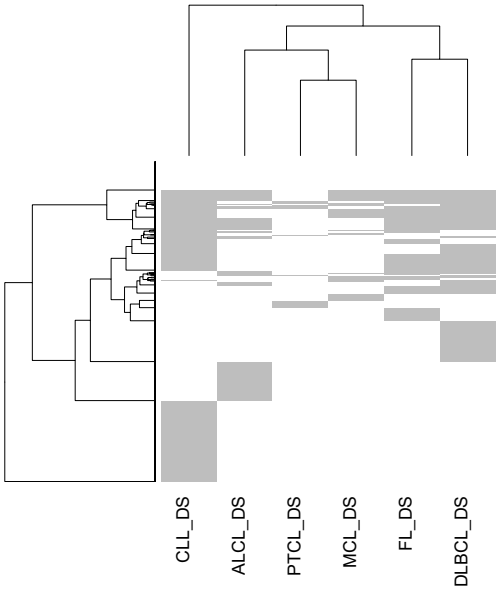


Figure 4.3: **Hierarchically clustered DS events per comparison.** Exon based differential splicing events (marked in grey) occurring in at least one subtype-control comparison are grouped by hierarchical clustering. While a considerable overlap for many subtypes is observed, some groups such as CLL, ALCL and DLBCL show also a substantial amount of specific splicing events. The highest number of DS events is observed for CLL (1395) while only few DS events are predicted for PTCL (337).

genes. It has been shown that down-regulation of keratin in cancer promotes migration and cell growth [Paccione et al., 2008, Fortier et al., 2013]. Consistent with this observation as well as withe the other results, functional analyses of the differentially spliced genes show typical characteristics of metastasis like cell adhesion.

4.3.5 SFs not Differentially Expressed but Differentially Central

Several SFs showed differential centrality in a number of conditions (Table 4.4), a few of them were also differentially expressed (Figure 4.4).

While some occur in most condition comparisons, amongst them DAZAP1, we also identified condition specific SFs such as RBM25 (see Table 4.4). For the most prominent candidates we screened literature to identify their known function as well as their involvement in disease. Amongst them are HNRNPM, RBFOX2, KHDRBS2, the ELAV-family as well as DAZAP1. For a detailed discussion of results see Section 4.4.1.

4.3.6 Who controls the Splicing Factors?

Splicing offers a fine-grained level of control in a cell. Yet, to understand the whole picture it is of substantial interest to investigate potential regulators who might impact

GO term	description
GO:0007398	ectoderm development
GO:0008544	epidermis development
GO:0030855	epithelial cell differentiation
GO:0060429	epithelium development
GO:0007155	cell adhesion
GO:0022610	biological adhesion
GO:0009913	epidermal cell differentiation
GO:0030216	keratinocyte differentiation
GO:0016337	cell-cell adhesion
GO:0042060	wound healing
GO:0022404	molting cycle process
GO:0022405	hair cycle process
GO:0001942	hair follicle development
GO:0031424	keratinization
GO:0042303	molting cycle
GO:0042633	hair cycle
GO:0031069	hair follicle morphogenesis
GO:0048730	epidermis morphogenesis
GO:0007160	cell-matrix adhesion
GO:0031589	cell-substrate adhesion

Table 4.3: Common **significantly enriched GO terms** in all subtypes related to differentially spliced genes.

the behaviour of splicing factors. To this end, we queried different resources, i.e. the database TRANSFAC [Wingender et al., 1996] and text mining results [Thomas et al., 2014] containing regulational information on transcription factors (TF) and their targets. Amongst these targets, we identified the SFs showing differential centrality in our approach.

For one SF detected as differentially central in our approach, PTBP2, we could identify a potential regulator in TRANSFAC. miR-133b, normally inhibiting cell proliferation, migration, invasion, and also promoting apoptosis, is a well known player in various cancer types. Due to its downregulation in cancer it performs an opposing effect on the cell [Zhao et al., 2014c, Zhao et al., 2013a, Hu et al., 2010, Xiang and Li, 2014, Novello et al., 2013].

For the subset of samples where miRNA sequencing data was available (for sample numbers see Table 4.5), we examined differences in expression of miR-133b between tonsil and the different subtypes. Table 4.5 shows a consistent downregulation for lymphoma in all comparisons. Due to a limited number of samples in the control group most comparisons show a p-value slightly over the significance limit while FC is constantly

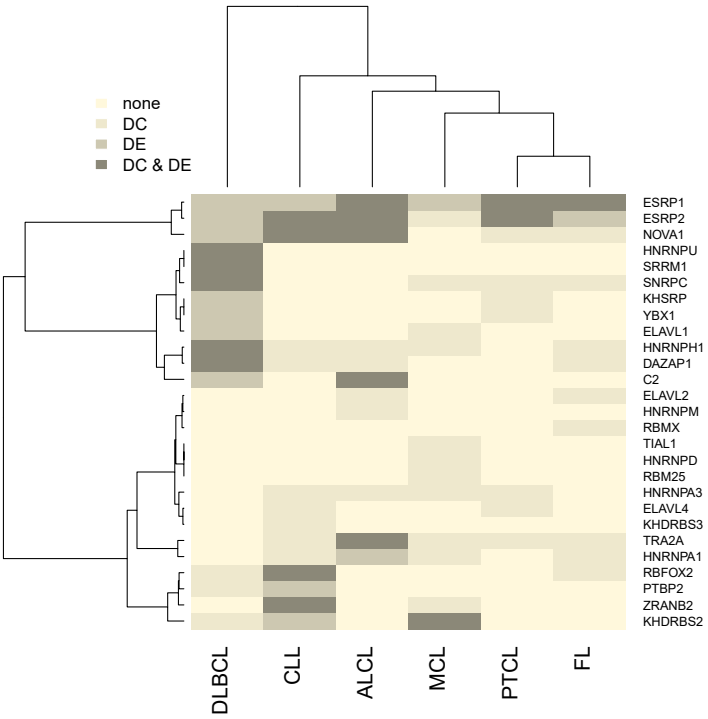


Figure 4.4: **Differentially central and differentially expressed splicing factors.** Hierarchical clustering of splicing factors based on their involvement on expressional and centrality based differences between control and lymphoma subtype. The heatmap includes all SFs being differentially central in at least one comparison. SFs can be either differentially central only (DC), differentially expressed but not differentially central (DE), differentially expressed and differentially central (DE and DC), or none.

high. This suggests a regulatory impact of miR-133b on PTBP2 in DLBCL (PTBP2 is DC) and CLL (PTBP2 is differentially expressed).

4.3.7 Differentially Central SFs and SNPs

Another important type of alteration which potentially changes the function of a protein and thus may lead to modified splicing activity of SFs are Single Nucleotide Polymorphisms (SNPs). Here we investigate the occurrence of SNPs in the SFs differentially central in lymphoma. We queried a large data set published by Klijn et al. [Klijn et al., 2014] comprising data from 675 human cancer cell lines. This data set also provides SNP information on various different cancer cell lines derived from different tissue types.

We first assessed whether the SFs differentially central in our data carried SNPs in lymphoma derived samples of the cancer cell line data set. We could identify SNPs for nine of our differentially central SFs; Table 4.6 shows them together with the number of SNPs in a group (lymphma / non lymphoma) normalized with the number of samples per group. Interestingly, a number of them have a higher SNP quantity (normalized to

entity	ALCL	CLL	DLBCL	FL	MCL	PTCL
TRA2A	+	+		+	+	+
HNRNPH1	+	+	+	+	+	
ESRP1	+	+		+		+
HNRNPA3	+	+			+	+
NOVA1	+	+		+		+
DAZAP1	+	+	+	+		
SNRPC			+	+	+	+
ESRP2	+				+	+
HNRNPA1		+		+	+	
RBFOX2		+	+	+		
ELAVL2	+			+		
ELAVL4		+				+
ZRANB2		+			+	
KHDRBS2			+		+	
HNRNPM	+					
C2	+					
KHDRBS3		+				
HNRNPU			+			
PTBP2			+			
SRRM1			+			
RBMX				+		
TIAL1					+	
HNRNPD					+	
RBM25					+	
ELAVL1					+	
KHSRP						+
YBX1						+

Table 4.4: **Splicing factors differentially central** in a certain lymphoma subtype compared to a tonsil control are marked with '+'.

the sample number per group) within the subgroup of lymphoma samples (see Table 4.6). To assess whether the SNP number of a certain gene is significantly higher inside the lymphoma group, we conducted a fisher exact test [Fisher, 1922] comparing the number of SNPs per gene in lymphoma samples to the number in all other cancerous samples.

Two SFs, HNRNPM (p=0.055) and HNRNPU (p=0.006) showed significant or close to significant p-values. Thus, we consider a lymphoma specific role for them, as with increased SNP number probability for an altered sequence and/or disease predisposition can increase.

entity	log2FoldChange	pval	Sample Number
ALCL	-5.5695	0.063238	9
CLL	-6.5652	0.032216	12
DLBCL	-2.4373	0.17737	39
FL	-3.4796	0.17426	20
MCL	-5.8855	0.078257	11
PTCL	-6.8632	0.065126	6
Tonsil			3

Table 4.5: **Mir-133b expression** comparison for tonsil versus each lymphoma subtype.

Gene	freq in lymphoma	freq in other cancers	ratio
HNRNPM	0.094	0.049	1.895
YBX1	0.019	0.004	4.295
KHSRP	0.038	0.028	1.357
HNRNPU	0.113	0.038	2.974
RBM25	0.019	0.032	0.586
ESRP1	0.038	0.088	0.430
DAZAP1	0.019	0.038	0.496
ZRANB2	0.038	0.038	0.991
RBFOX2	0.038	0.067	0.560

Table 4.6: **Frequency of SNPs per gene in lymphoma and non-lymphoma cancer cell lines.** The number of SNPs per gene is normalized by group size.

4.4 Discussion

Our goal is to identify splicing factors responsible for aberrant splicing in lymphoma. In this work, we present a method for identification of such potential regulators using only transcriptome data. Instead of looking at expression changes only, we use an indirect measure, which allows to encompass changes on different levels. Specifically, we integrated differentially spliced exons with the set of known splicing factors in condition-dependent correlation networks. Those SFs were considered regulatory candidates, which showed differential centrality between a diseased and a control network.

4.4.1 Biological Assessment of Results

Comparing differentially expressed SFs to those differentially central reveals a rather consistent picture of the ongoing changes in several lymphoma subtypes. Three main motifs, EMT-alike changes, MEK/ERK pathway related changes as well as neuro-oncological associated changes are observed in both, DE and DC results.

EMT, an important mechanism in regular development including altered cell-cell adhesion and cell migration, is often hijacked in cancer and thus leads to metastasis and reduced susceptibility to apoptosis. EMT related changes are known for different lymphoma subtypes as, for instance, mantle cell [Sanchez-Tillo et al., 2014] and diffuse large B-cell lymphoma [Lemma et al., 2013]. Several additional cancer types such as breast cancer [Xu et al., 2014, Shapiro et al., 2011, Horiguchi et al., 2011], somatotroph adenoma [Lekva et al., 2013] and colon cancer [Venables et al., 2013] exhibit this mechanism. An epithelial to mesenchymal isoform switch is observed in clear cell renal cancer [Zhao et al., 2013b] and for pancreatic cancer, ESRP1 is a potential prognostic factor [Ueda et al., 2013].

HNRNPM is known to play a crucial role in splicing-related breast cancer metastasis [Xu et al., 2014]. By increasing TGFβ signaling, HNRNPM impacts the splicing of CD44 through competing for cis-regulatory binding sites with ESRP1. Current research also suggests a role in colorectal cancer [Chen et al., 2014]. In this condition, HNRNPM dependent metastasis and cancer recurrence is observed. RBFOX2 is an important driver of mesenchymal tissue specific splicing in normal and cancer tissue and splicing differences between normal and tumor are similar to those between mesenchymal and epithelial [Venables et al., 2013]. KHDRBS2 (alias:SLM1, SAM68) is know to influence a certain CD44 isoform abundance. More specifically, KHDRBS2 promotes the inclusion of exon 5 in CD44 due to ERK-mediated phosphorylation [Matter et al., 2002].

An interesting validation approach based on our investigation of the SNP frequency in differentially central SFs would be the sequencing of HNRNPM and HNRNPU in our lymphoma samples. A high number of SNPs might lead to altered function, especially to altered interaction patterns with other (splicing-relevant) proteins such as SFs or histones, and thus to a potential involvement in lymphoma-related aberrant splicing.

Also, as our analyses suggest DS in ESRP1/2 in some subtypes, a more detailed, experimental analysis and quantification of the expression status of the two could reveal the presence of different isoforms.

The second group of changes can be summarized as **MEK/ERK pathway** related. An important target of cancer related studies and a high ranked SF in our approach is HNRNPH1. In colorectal cancer HNRNPH expression showed significant association with tumor stage, i.e. with metastasis as well as with survival [Hope and Murray, 2011]. In Gliomas, HNRNPH acts as an oncogenic splicing switch by promoting aberrant isoforms leading to resistance to apoptosis and metastatic tendencies [LeFave et al., 2011]. The expression of A-RAF, a protein involved in the MEK-ERK signaling pathway, is controlled by HNRNPH. Suppressing of either A-RAF or HNRNPH leads to apoptosis while an upregulation promotes resistance to apoptosis [Rauch et al., 2010]. Not only HNRNPH but also DAZAP1 is known to act in this pathway. A recent publication identifies DAZAP1 as a MEK/ERK pathway influenced splicing regulator of cell proliferation and migration [Choudhury et al., 2014]. The MEK/ERK signaling pathway controls the activity of DAZAP1 by phosphorylation, raising the question of a defect in the phosphorylation cascade. To this end, we examined the expression of all genes involved in the pathway, which revealed differential expression of pathway members for ALCL (MRAS, MYC) and CLL (RRAS, MAPK1/ERK2, RAF1/CRAF, BRAF and

MAP2K1/MEK1). One condition, DLBCL, showed no differential expression of members of the pathway, but of DAZAP1 itself. In summary, we see substantial expression changes in the MEK/ERK pathway which might lead to an alteration of the function of DAZAP1 and thus to altered splicing. In this context, the downregulation of miR-133b in all lymphoma subtypes is interesting, as a recent publication attributes this miRNA a role in ERK-signalling [Feng et al., 2013].

Neuro-oncological changes in our data comprise, for instance, ELAVL2, a neuron specific RNA binding protein and tumor antigen. ELAVL2 is associated with small-cell lung cancer [D'Alessandro et al., 2008,Kazarian and Laird-Offringa, 2011], where mRNA levels were significantly raised in blood, as well as in renal carcinoma [Stickel et al., 2009]. Not only ELAVL2, but also ELAVL1 and ELAVL4 appear amongst the differentially central SFs. The ELAV-family is an important post-transcriptional regulator for NOVA1 and impacts the splicing activity of the latter [Ratti et al., 2008].

4.4.2 Network-based Differential Centrality

It is currently technically impossible to capture all changes in a cell leading to aberrant physiological circumstances. Nevertheless, these alteration influence each other and might thus be detected indirectly by investigating, as in our example, expression via networks.

Several reasons favor the differential network approach to DS. First, exons and SFs are known to have a many-to-many association, meaning one exon can be influenced by many SFs whereas one SF can influence several exons. Second, networks enable the discovery of potentially functional modules, such as pathways, which might be disturbed.

A more general argument towards networks is the application to cancer data itself. On the one hand, this phenotype is a complex disease who's emergence is based on multiple malfunctions, on the other hand it might also have very heterogeneous causes. An observed phenotype can, for instance, be provoked by different defects in a functional unit, finally leading to the same effect but being caused by different alterations.

For prioritization of entities in networks a variety of *centrality measures* have been used [Koschützki and Schreiber, 2008]. Depending on the aim of the study, different measures are of favor. Detection of essential proteins in protein-protein interaction networks is better achieved by measures based on interconnectivity such as the clustering coefficient [Estrada, 2006]. For regulatory networks, on the other hand, betweenness centrality is preferable [Yu et al., 2007], as so-called bottleneck proteins are more important for information flow than hub proteins. Our approach thus favors betweenness centrality.

4.4.3 Correlation Cutoff and Result Stability

Our interest in the correlation cutoff dependence of results motivated us to vary the latter in steps over the whole spectrum. As the smallest group in our setting is the control group (comprising 9 samples), we used a statistical test for Pearson correlation to determine the significant correlation cutoff r, which is 0.67 for our control group. We compared the top ten differentially central SFs for networks (ALCL vs tonsil) based on

this cutoff to the ones with varying cutoffs to assess whether our results are dependent on a - potentially arbitrary - cutoff. Eight out of the ten SFs occur in both, the significant as well as the 'majority vote' cutoff scenario. For our analysis we used the majority vote, as we deemed it more robust.

4.4.4 Evaluation

Little is known about splicing and lymphoma. Therefore, no 'gold standard' of spliced entities or splicing-regulatory trans-acting factors exists. We thus evaluated our approach by assessing the enrichment of differentially expressed SFs amongst our differentially central candidates, arguing that these are the most reliably involved regulators and should thus be highly ranked in our approach. Two (ALCL, DLBCL) out of three conditions we considered adequate for our purpose showed an enrichment of differentially expressed SFs amongst the top differentially central ones. The third condition, CLL, contains the highest number of differentially spliced exons as well as differentially expressed SFs. Potentially, the number of involved SFs is thus higher and it is therefore harder for differentially expressed SFs to rank under the top positioned ones. As we apply a rather rigorous cutoff by deeming only the SFs occurring in at least half of the cutoff settings as relevant, we face the situation that a growing number of differentially expressed SFs will reduce significance. However, this method aims at comparisons where little differential expression, and thus little direct evidence for potential causes is given.

4.5 Conclusions

Most of the identified SFs are well known in the context of cancer. While this is a valuable support for the reliability of our results, the truly interesting results are those which lead to candidates undiscovered to date. Interestingly, DAZAP1, a SF showing DC in most subtypes (and DE in two of them) was only recently set into context with MEK/ERK pathway which is connected to CD44 splicing [Choudhury et al., 2014]. A new candidate for regulation of splicing in lymphoma based on our results is TRA2A. While being differentially central in most conditions (as well as differentially expressed in one of them), little to nothing is known about the role of TRA2A in the context of cancer or other diseases. This highly ranked SF might thus be an eligible candidate for further investigations. Based on the SNP frequency in the cancer cell data set queried, the SNP frequency of HNRNPM in ALCL samples as well as HNRNPU in DLBCL samples poses a promising investigation target. Interestingly, HNRNPU is also differentially expressed in DLBCL. Overall, the probably most beneficial aspect of our method is the fact that alterations in our regulatory candidates are not necessarily restricted to the transcriptomic level, while only using transcriptomic data.

5 Multi-level Comparison of Exon Array and RNA Sequencing Data

For the last two decades, microarrays have been the indisputable method of choice to quantify the transcriptome in high-throughput manner. However, during the past decade, RNA sequencing techniques are gradually complementing and replacing microarrays [Wang et al., 2009b, Zhao et al., 2014b]. To preserve microarray-based knowledge, a thorough evaluation of the comparability of the results produced by the different technologies is indispensable. Several studies pursue this task on the gene level, reporting high concordance on on present/absent entities, correlation of expression values [Raghavachari et al., 2012], fold change direction [Bottomly et al., 2011] and differentially expressed genes [Wang et al., 2014a, Wang et al., 2014b].

Generally, RNA sequencing is found to complement and extend the merits of microarrays [Malone and Oliver, 2011]. This is shown by higher verification rates [Nault et al., 2015, Wang et al., 2014a, Wang et al., 2014b] of RNA-Seq results as well as the widely observed finding of higher sensitivity in the lower expression ranges [Bottomly et al., 2011, Wang et al., 2014a, Wang et al., 2014b].

Nevertheless, splicing microarrays, such as exon arrays or junction arrays, provide the possibility to study splicing alterations, allowing for comparing the latter one to RNA sequencing results with respect to differences in detected splicing events. In contrast to studies on the gene level, these comparisons are rare and often lack explanatory power due to small sample sizes which impede statistical tests, rarely used types of microarrays, or a result set too small for representative comparisons.

Irrespective of the technology, two general approaches exist for assessing alterations in splicing. One possibility is based on the isoform level. Hereby, the expression of each transcript is quantified separately and subsequently compared to the respective expression intensity in the control group. This approach requires knowledge on the transcripts, which either excludes unknown ones or operates with (rather heuristic) transcript prediction algorithms. An alternative is the assessment of differential exon usage (see Section 3.1 and 3.2). Here, the individual exon expression is usually set into relation to the gene expression intensity. While it is possible to use isoform expression intensities instead, the gene level normalization approach provides a more neutral solution, omitting the task of isoform level quantification.

Very few studies examine the comparability of performance on the exon level [Raghavachari et al., 2012, Xu et al., 2011]. Raghavachari et al. [Raghavachari et al., 2012] analyze the concordance of Affymetrix Exon Arrays and Illumina sequencing. While gene level results are meaningful due to a high number of differentially expressed genes, a lower number of genes displays differential splicing (1 vs 16 without overlap) which impedes

a robust comparison. A second study by Xu et al. [Xu et al., 2011] is based on a rather rarely used microarray (Glue Grant Human Transcriptome Array); GEO [Edgar et al., 2002] reports only 10 studies based on this technology. Bradford et al. [Bradford et al., 2010] compare the Affymetrix Exon Array to ABI Solid, using only a single sample for the actual platform comparison. Subsequently, only comparisons in fold change rather than statistical tests are applicable. The most interesting comparison is published by Dapas et al. [Dapas et al., 2016], not only comparing differentially expressed genes but also differentially expressed isoforms based on RNA sequencing and exon arrays. Correlation of expression and fold change for genes and isoforms reveal a higher agreement for gene comparison. Indeed, isoform quantification is still a hard problem, and various tools lead to substantially varying results [Dapas et al., 2016]. Even the application of a multitude of quantification tools does not exceed a maximal correlation coefficient of 0.5 [Dapas et al., 2016]. Thus, a comparison on the exon level, avoiding the supposedly biasing step of isoform quantification might yield a clearer picture of the ongoing changes on the splicing level.

In this chapter, we aim at assessing the comparability of differential exon usage detected from exon arrays and RNA sequencing data. To this end, we developed a multi-level framework, enabling comparison of both technologies not only on the level of differential splicing, but on all antecedent levels. We provide insight on (1) the most basic probe level, by implementing a maximally comparable data pre-processing procedure, (2) we enable comparison of the next-higher aggregation step for probe set/exon data as well as (3) a gene level comparative approach. Finally, we (4) quantify differential splicing based on several methods, one which is typically used for the respective technology and others which are designed for both technologies. By applying the same DS detection method to both data types, we aim at reducing method-induced bias, which can be rather high even for different methods applied to the same technology.

The comparison on several levels enables a detailed understanding of the data characteristics from each technology. Only by this thorough multi-step evaluation, we can fully understand to what extent the two technologies are comparable and what the limitations are. We apply our approach to two cancer data sets.

5.1 Levels of Comparison for Exon Array and RNA Sequencing Data

We first give a detailed insight on the different comparison levels implemented in our framework. We describe pre-processing steps as well as methods used for the acquisition of differential expression on various levels together with the mode of comparison applied on the individual level. For an overview see Figure 5.1.

5.1.1 Probe Level Comparison

The most basic level of comparison is posed by the bare sequence responsible for quantification of expression. For exon arrays, this is the probe level, thus we refer to this

comparison step as the *probe level comparison* in the following. For RNA sequencing, the genomic positions corresponding to the probe locations are selected. Several aspects of this step need to be considered to ensure comparability of expression values. First, genomic positions of the probes have to be assessed based on HG19, as Affymetrix provides probes positions based on HG16. We therefore used BLAST to map probe sequences to HG19 and determine the positions of uniquely mapping probes.

Further, the corresponding count values for the sequencing data have to be assessed based on their genomic positions in HG19, so only reads actually covering the probe sequence are used. Note that this step ensures the maximally possible comparability of data as both technologies should measure the exact same expression. All other levels are less comparable in terms of the genomic input positions, as exon array data is an aggregate of individual probes, while RNA sequencing data covers the whole region of the entity under consideration in terms of expression.

Due to differing properties of the two technologies, such as diverging data scales, filtering of the expression data can be necessary.

The pre-processed and filtered data is then compared for matching samples, giving an overview on the correspondence of the expression data on the most downstream level.

5.1.2 Probe Set Level Comparison

The second level of comparison is based on the so-called probe sets defined on the exon array, which group probes corresponding to a certain feature together. Probe sets represent exons of a gene. The equivalent expression values for the sequencing data are exon counts. Note that even in the case of an 'injective' mapping, i.e. a probe set is mapped to only one exon, the probe set value is based on the values of its probes only, and does therefore - in most of the cases - not cover the whole exon sequence. Probe sets, i.e. their corresponding probes, are designed to reduce cross-hybridization with other genes, nevertheless, cross-mapping with exons within a gene is possible. To maximize comparability of the used expression values for this level, we filter the exon array data for injective mapping hits by using only probe sets that match to exactly one exon.

This pre-processed data is then used to assess correlation between corresponding samples as well as correlation of exon fold changes between groups.

5.1.3 Gene Level Comparison

The next-higher level is the gene level. To this end, data is summarized based on probe sets in the exon array case. For sequencing data, counts on gene level are assessed. Subsequently, expression data from both technologies is analyzed for differential gene expression using state of the art analyses. In the case of array data, a linear model (limma [Ritchie et al., 2015]) is applied, for sequencing data DESeq2 [Love et al., 2014] is used to determine genes significantly differing in expression between the two groups.

Using these results, the correlation of fold changes gives an overview on the comparability of the data. Furthermore, the overlaps between significantly differentially expressed genes are quantified and tested for significance.

5.1.4 Comparison of Differential Splicing

Combining gene level and probe set level data enables the final step of analysis, the investigation of differential splicing. In contrast to differential gene expression, this analysis is less established and poses more challenges (see also Chapter 3.1). In consequence, methods analyzing differential splicing tend to produce rather heterogeneous output. To avoid that this bias influences our results, we use two methods applicable to data from both exon arrays and RNA sequencing. These two, ARH and SplicingCompass, have been introduced in Chapter 3.2. For further comparison, we also apply DEXSeq [Anders et al., 2012], a method widely used for the detection of differential exon usage in RNA sequencing data.

Results are then compared by determining the overlaps between different methods and technologies as well as their significance.

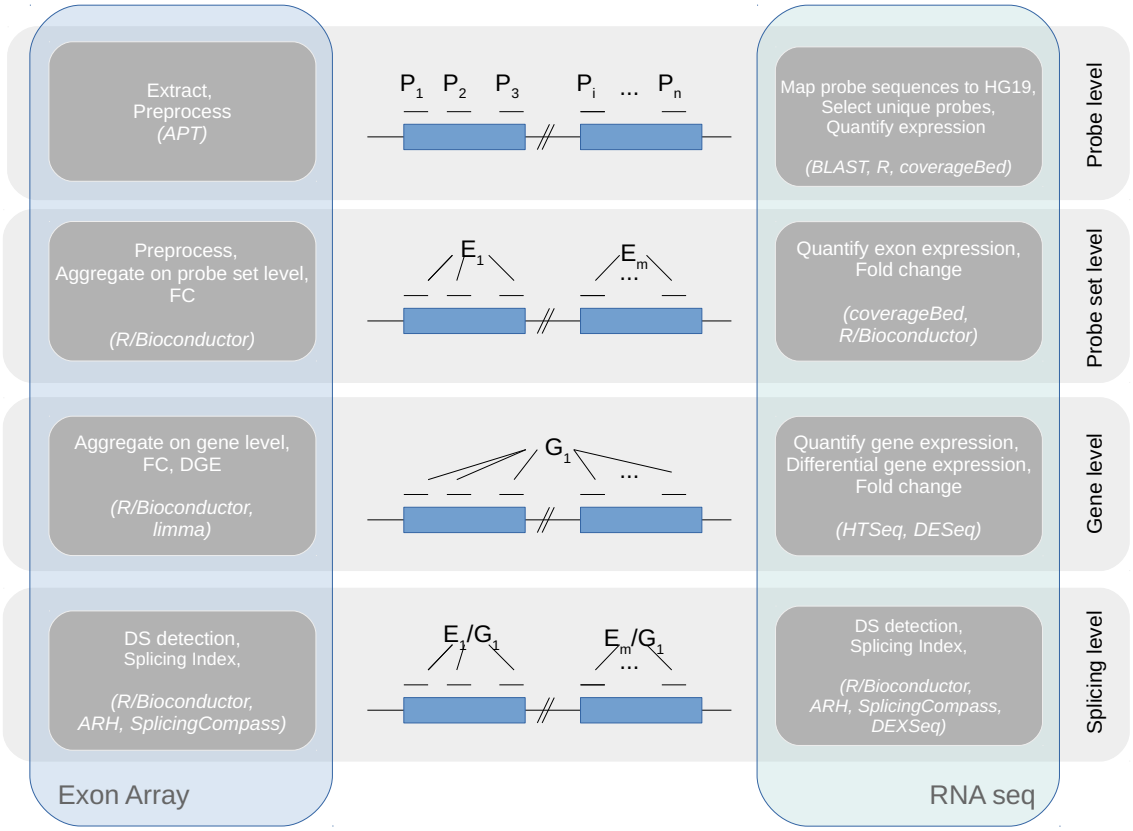


Figure 5.1: **Multi-level framework for the assessment of comparability of exon array and RNA sequencing data with respect to different analyses.** Each of the four levels includes different expression sets as well as the corresponding analysis strategies and tools.

5.2 Multi-Level Preprocessing and Analysis

This section describes the data used as well as the pre-processing and analysis steps taken. Besides the tools applied, references and annotation used for the analyses are introduced.

5.2.1 Expression Data

For the comparison of exon array data to RNA sequencing data with respect to the comparability of differential splicing detection, RNA sequencing data for a subset of unmatched biological samples of the earlier used exon array cohort (see Chapter 2.3) was generated. Six samples, three DLBCL and three Tonsil samples, were analyzed in both technologies, enabling direct comparison. Note that due to the generation of the sequencing data at a later time point, the RNA used is not identical to the one used for the exon arrays, although it originates from the same biological sample. Tumor heterogeneity, for instance, might lead to differences, which can not be quantified by this approach. In the following, we will refer to the samples as Tonsil-1 to Tonsil-3 and DLBCL-4 to DLBCL-6 respectively. We discuss results based on this data in great detail. Additionally, we also applied our multilevel-framework to a second data set comprising four glioblastoma multiforme and four control samples. Results for this data set are presented and contrasted with results for the DLBCL data set in Chapter 5.3.7.

Samples	No. initial reads	initial GC%	No. trimmed reads
Tonsil-1	75,648,251	52	75,074,381
Tonsil-2	87,542,875	53	87,143,795
Tonsil-3	71,646,252	51	70,994,964
DLBCL-4	85,940,819	51	85,568,553
DLBCL-5	85,464,082	51	84,371,002
DLBCL-6	82,667,161	51	82,381,667

Table 5.1: **Result summary of fastQC and trimming.**

For sequencing, an Illumina Hi-Seq machine was used, producing single reads of 100 base pairs. An overview on the number of reads before and after trimming with Trimmomatic [Bolger et al., 2014] is given in Table 5.1. Reads were quality controledl using FastQC [Andrews et al., 2010] and mapped to the genome (HG19 and HG38) with STAR [Dobin et al., 2013], an alignment tool allowing for spliced alignment. A result summary of the alignment is shown in Table 5.2, indicating a good overall output.

Expression matrices for the different comparison levels. To allow for comparison of the expression data from both technologies on the defined levels, each of them requires its corresponding expression set. In the following, we describe generation of the data used on the four levels.

Preprocessing of the exon array samples is accomplished as described in Section 4.2.1, except for the probe level data. For acquisition of the latter see the following paragraph.

Samples	input reads	uniquely mapped	mapped to multiple loci	unmapped	Annotated (sjdb) splices	Non-canonical splices
Tonsil-1	75,074,381	90.52 %	7.89 %	1.54 %	22,426,796	53,679
Tonsil-2	87,143,795	90.87 %	7.60 %	1.50 %	28,043,725	60,345
Tonsil-3	70,994,964	91.18 %	7.65 %	1.12 %	21,909,709	69,769
DLBCL-4	85,568,553	92.47 %	7.13 %	0.37 %	33,100,432	79,533
DLBCL-5	84,371,002	93.29 %	6.27 %	0.42 %	32,892,221	46,350
DLBCL-6	82,381,667	91.85 %	7.71 %	0.42 %	35,950,647	55,121

Table 5.2: Result summary of read alignment using STAR.

1. **Probe level.** Probe level data from exon arrays was extracted and quantile normalized with Affymetrix Power Tools [Lockstone, 2011]. To allow reasonable comparison with sequencing data, only read counts mapping to the actual probe sequences were extracted from RNA-Seq data. Thus, the genomic positions of the probes, designed based on HG16, were assessed by mapping all probe sequences to HG19 using BLAST [Altschul et al., 1990], and retaining only uniquely mapping probes. Based on the probe positions in HG19, counts were assessed with coverageBed, a tool from the genomic analysis suit Bedtools [Quinlan and Hall, 2010]. Both data sets were log2-transformed prior to comparison.

2. **Probe set level.** Data for the probe set level (i.e. exon level) was compared based on Ensembl exon ids. For exon arrays, data is preprocessed and aggregated as described in [Rodrigo-Domingo et al., 2013]. Subsequently, the resulting probe sets were mapped to Ensembl exon ids using biomaRt [Durinck et al., 2009]. An extensive filtering was applied, to ensure injectivity of the resulting mapping. Finally, probe sets mapping to the same Ensembl id were aggregated by using their mean. For the sequencing data, RPKM counts (reads per kilobase of transcript per million mapped reads) per exons were used.

3. **Gene level.** Gene expression values for exon array data were generated by a further aggregation step as described in [Rodrigo-Domingo et al., 2013]. For sequencing data, the corresponding gene transfer format (GTF) annotation file from Gencode [Harrow et al., 2012] was used to generate gene level count data with HTSeq [Anders et al., 2014] as Gencode is preferable in the context of splicing [Frankish et al., 2015].

4. **Splicing level.** Splicing level data is based on the probe set level, i.e. exon level data. Additionally, the information of the corresponding gene for each exon is used as input for the different tools. Most tools for sequencing data require specific input data. SplicingCompass, for instance, uses CCDS (Consensus coding DNA sequence) data and requires raw counts from coverageBed, DEXSeq [Anders et al.,

2012] comes with its own counting script. ARH uses the exon-level expression quantification.

5.2.2 Analysis and Comparison Methods for the Different Comparison Levels

For the two most basic levels, probe and probe set level, we used Pearson correlation as a method of similarity acquisition.

1. On the **probe level**, we correlated expression values of the corresponding samples.

2. On the **probe set level**, in addition to correlating expression of corresponding samples, we used the fold changes per probe set / exon between groups (DLBCL vs. Tonsil) as correlation input.

3. On **gene level**, we compared the set of differentially expressed genes as well as all genes in terms of overlap and their fold change correlation. We thus assess differentially expressed genes by using state of the art methods for each technology. For exon array data, limma/Bioconductor [Ritchie et al., 2015] was applied, i.e. differential expression is determined using a linear model. Sequencing data was analyzed with DESeq2 [Love et al., 2014]. Here, the variance-mean dependence is estimated and data is tested for differential expression based on a model using the negative binomial distribution.

4. For the **splicing level** we expect differences due to technology. Therefore, we wanted to minimize bias introduced by different methods and chose to use two methods, ARH [Rasche and Herwig, 2010,Rasche et al., 2014] and SplicingCompass [Aschoff et al., 2013], applicable to both data types for comparison. Additionally, we use a widely applied differential exon usage detection method, DEXSeq [Anders et al., 2012] for an inner-technological reference.

Our main genome version for comparison is HG19, i.e., GRCh37.p13 from Gencode. Yet, we reran our RNA sequencing analyses with GRCh38, to also provide insight on differences induced by the useage of the current genome version.

5.3 Results

The following section presents the results of the comparisons on the four different levels. Note that each level includes its own filtering steps which are not influenced by other levels. Thus, probes excluded on the probe level might be part of the gene level and vice versa. The idea motivating this procedure is twofold. First, we would like to choose the most unambiguous entities on every level, without, second, loosing too much of the data along the way, i.e. the levels. While this work can be used to select filter criteria for better result concordance of exon array and sequencing data, we belief that the most important application scenario is a comparison of current, sequencing-based results to

older, public exon array results. Thus, before proposing filter criteria for better result concordance, we want to assess comparability of the technologies as they are used up to date. The stricter the filtering criteria, the more coverage of relevant genomic regions is lost.

5.3.1 Probe Level

For a first overview on the data, we compared the distributions of log2 transformed data from both technologies as displayed in Figure 5.2. While the count data experienced no further transformation than the logarithm (except for a pseudo-count of 1 to avoid problems in logarithmization), exon array data was additionally quantile normalized. While RNA sequencing is designed to resolve expression of entities on a single transcript level, quantification with exon arrays returns an approximation of expression. Therefor, scales in Figure 5.2 are not directly comparable. Nevertheless, they illustrate the expected divergence in size of the represented expression ranges, which is much broader for RNA sequencing data. The boxplot also illustrates the better expression resolution for lower expression ranges in RNA sequencing data, the second quartile, i.e. the lower half of the box, spans a higher range (relative to the third quartile) for RNA sequencing data compared to exon array data.

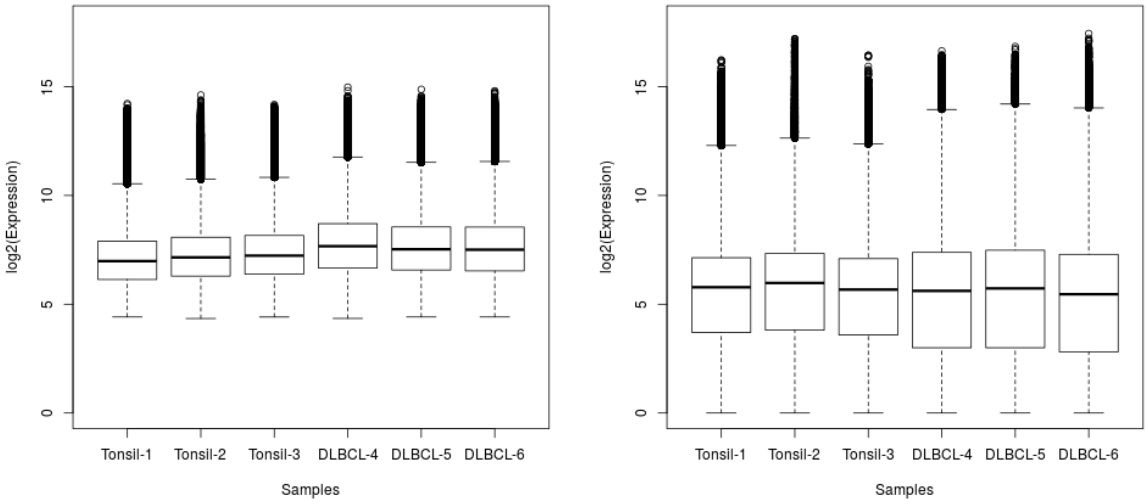
In order to determine the genomic categories of the data points, we intersected the genomic locations of the probe sequences with the positions of the exons and genes in our annotation gtf (see Table 5.3). From the 1,082,385 core probes used, 961,488 were uniquely mapped to the genome. Of those, 894,029 were located in consensus coding sequence (CCDS) exons and 957,950 were found in regions defined by gene coordinates. Thus, 63,921 core probes map to intronic regions or non-CCDS exons, and 3,538 probes mapped outside of annotated genes.

	number of probes	% core	% unique
core probes used	1,082,385	100%	
uniquely mapping probes	961,488	88.8%	100%
probes in exons	894,029	82.6%	93%
probes in genes	957,950	88.5%	99.6%
probes in introns or non CCDS exons	63,921	5.9%	6.7%

Table 5.3: **BLAST results compared to HG19 annotation.** The positions of the probes were intersected with annotated entities.

Based on these preprocessing steps, we selected the most reliable probes for comparison, i.e. only uniquely mapping probes are selected.

The filtered values were then used to access a sample-specific correlation of expression values, comparing each sample in the expression set of one technology to the corresponding sample in the expression set of the other technology. The sample-wise scatter plots in Figure 5.3 give an overview on the Pearson correlation ranging from $r = 0.11$



(a) Exon array data

(b) RNA sequencing data

Figure 5.2: **Boxplot of log2 probe level expression intensities for exon array and RNA sequencing data.** Only values of probes uniquely mapped to the genome are considered. Exon array data was quantile normalized and core probes were extracted before filtering for uniquely mapping probes.

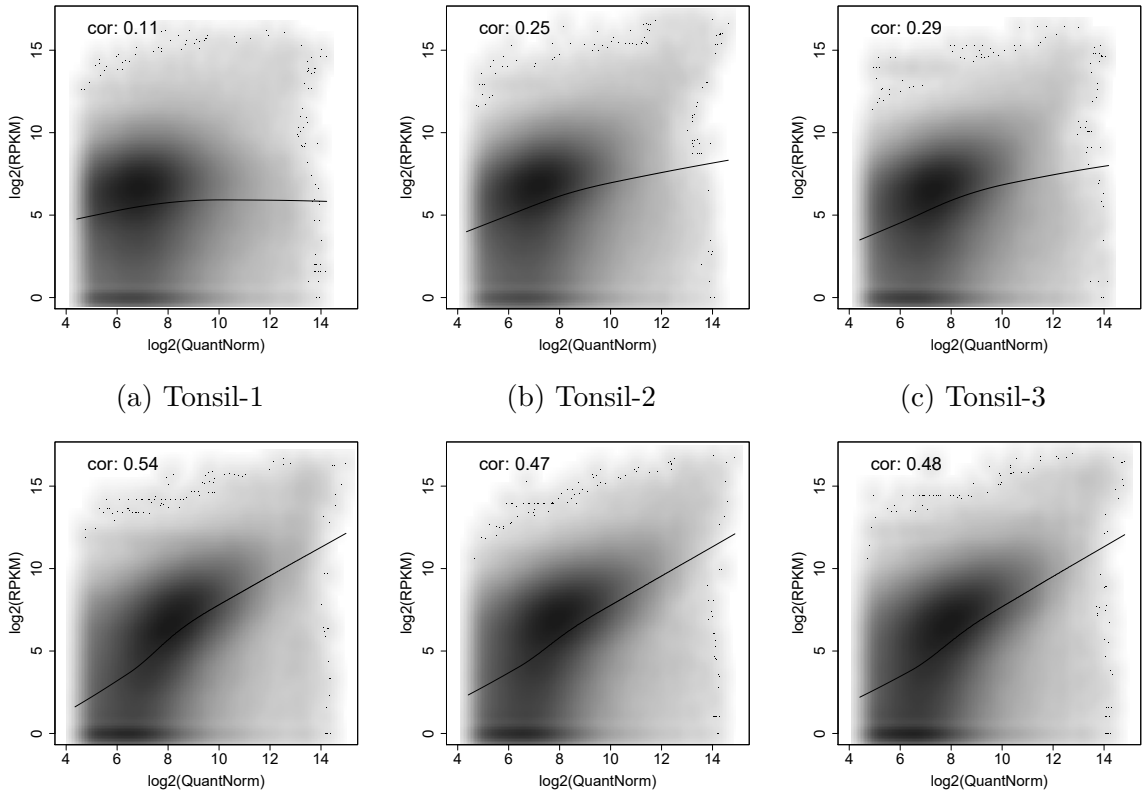
to $r = 0.54$. Each correlation was highly significant. Correlation is significantly higher ($p = 0.05$, Wilcoxon-test) for DLBCL samples compared to Tonsil.

Figure 5.4 shows the distribution of the expression values compared per sample. In concordance with Figure 5.2, RNA sequencing data is characterized by a high and widespread amount of low expression data points, while exon array data is condensed on a small expression range with rather normally distributed values.

Last, we visualized expression of all probes for two selected genes to give an impression on the corresponding level-values as well as to display a gene exhibiting differential splicing in comparison to one which does not. To this end, we chose DC44 since it is known to be differentially spliced in lymphoma as well as DRAM2 which was randomly selected from the set of genes not predicted as differentially spliced by none of the methods applied in this approach. Figure 5.5 visualizes the two genes on the probe level. Sequencing data exhibits rather stable probe level expression, while exon array data shows more variation. Despite expectations, no signs of differential splicing are visible on this level.

5.3.2 Probe Set Level

Probeset level filtering was done independently from the previous level. We started with all 284,805 core probe sets in our data and excluded all probe sets containing cross-hybridizing probes [Rodrigo-Domingo et al., 2013]. This filtering step reduced our set



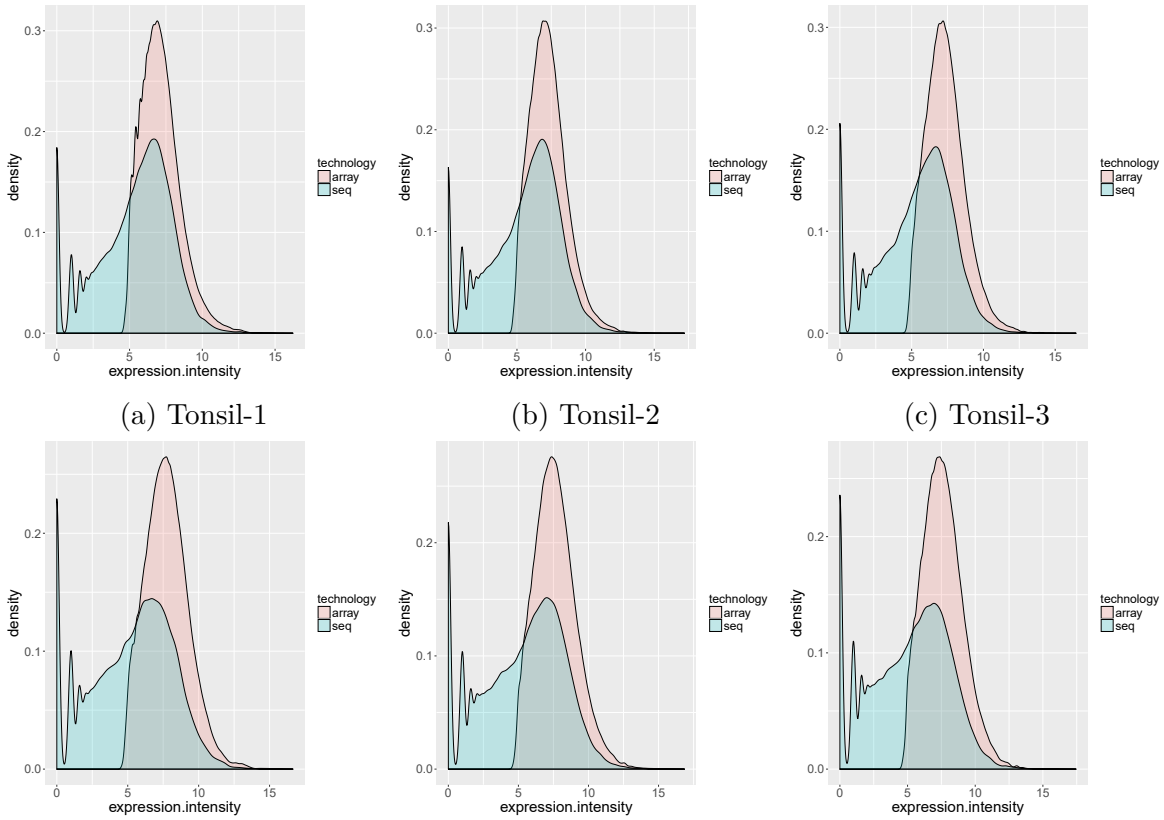
(a) Tonsil-1 (b) Tonsil-2 (c) Tonsil-3

(d) DLBCL-4 (e) DLBCL-5 (f) DLBCL-6

Figure 5.3: **Probe level scatter plot of sample-wise correlations.** The corresponding expression values from both technologies are compared per sample. Only probe level values of uniquely mapped probes are incorporated. A lowess line indicating locally weighted regression is included in every plot for a better visualization of comparability. All correlations are highly significant.

to 229,023 probe sets. Next, we mapped all remaining probe sets to Ensembl exon ids using Biomart. From this mapping, we extracted all injectively mapping probe sets (3,887 corresponding to 1,143 exons), and aggregated their expression values using the mean. For the sequencing data, we extracted the corresponding data based on the ensembl exon ids, using RPKM values.

The sample by sample correlation improved for all Tonsil samples with respect to the probe level correlation ($r = 0.2, r = 0.3, r = 0.3$), but deteriorated slightly in the group of DLBCLs ($r = 0.5, r = 0.44, r = 0.44$). All correlations were highly significant. Figure 5.6 shows scatter plots of all sample wise comparisons and indicates the respective correlation values. As on the probe level, correlations are superior for the DLBCL samples with an even lower p-value ($p = 0.04$). Also, the ranks on correlations of the samples are stable between comparison levels, i.e. relative comparability is coherent over levels.



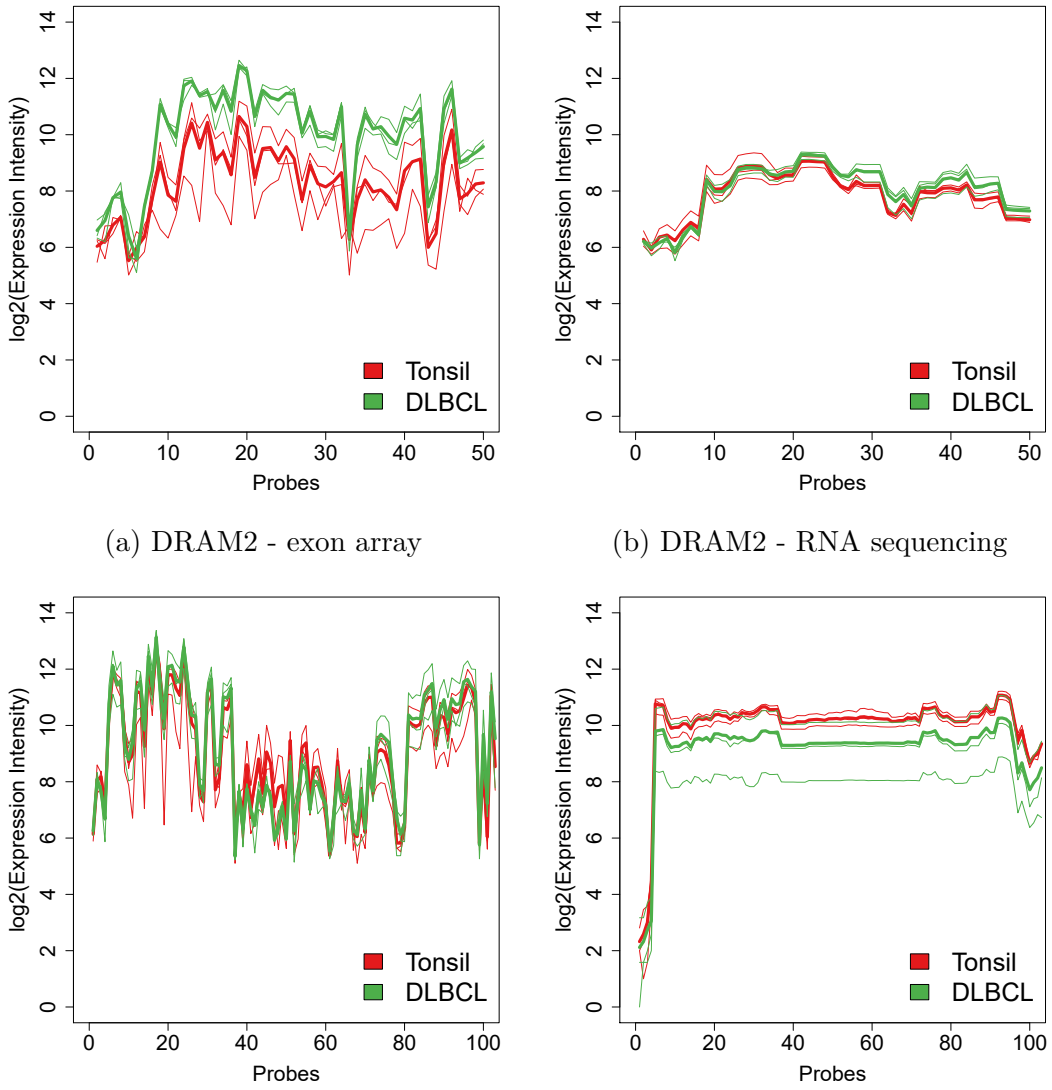
(a) Tonsil-1 (b) Tonsil-2 (c) Tonsil-3

(d) DLBCL-4 (e) DLBCL-5 (f) DLBCL-6

Figure 5.4: **Histogram of probe level expression for all samples and both technologies** based on uniquely mapped probes. Red denotes exon array data, blue encodes RNA sequencing data.

To gain more group-specific data insight, we assessed the fold changes between DLBCL and Tonsil samples for each technology based on our filtered data and correlated the two. Figure 5.7 shows the result, a fold change correlation of $r = 0.19 (p = 1.9533e - 07)$. Sample level probe set correlation (Figure 5.6) shows a high amount of data points with low to moderate expression values for array data, which have zero or low values in sequencing data. This specificity is even more clearly notable for the probe level distributions in Figure 5.4. To test whether these exon array values can be deemed as noise and are thus the cause of the rather low fold change correlation, we repeated the analysis with only that 50% of the data points, which had the highest expression in the exon arrays. This procedure improves correlation to $r = 0.29$ $(p = 7.943459e - 09)$. Figure 5.7b shows the same filtering results based on RNA sequencing data. Here, FC correlation improves to $r = 05$.

On probe set level, the two example genes are compared based on probe sets for exon array data and based on exons for sequencing data displayed in Figure 5.8. Both are sorted based on genomic coordinates. The picture on this level is clearer in terms of splicing, CD44 shows a region of separation between the group mean values, while



(a) DRAM2 - exon array

(b) DRAM2 - RNA sequencing

(c) CD44 - exon array

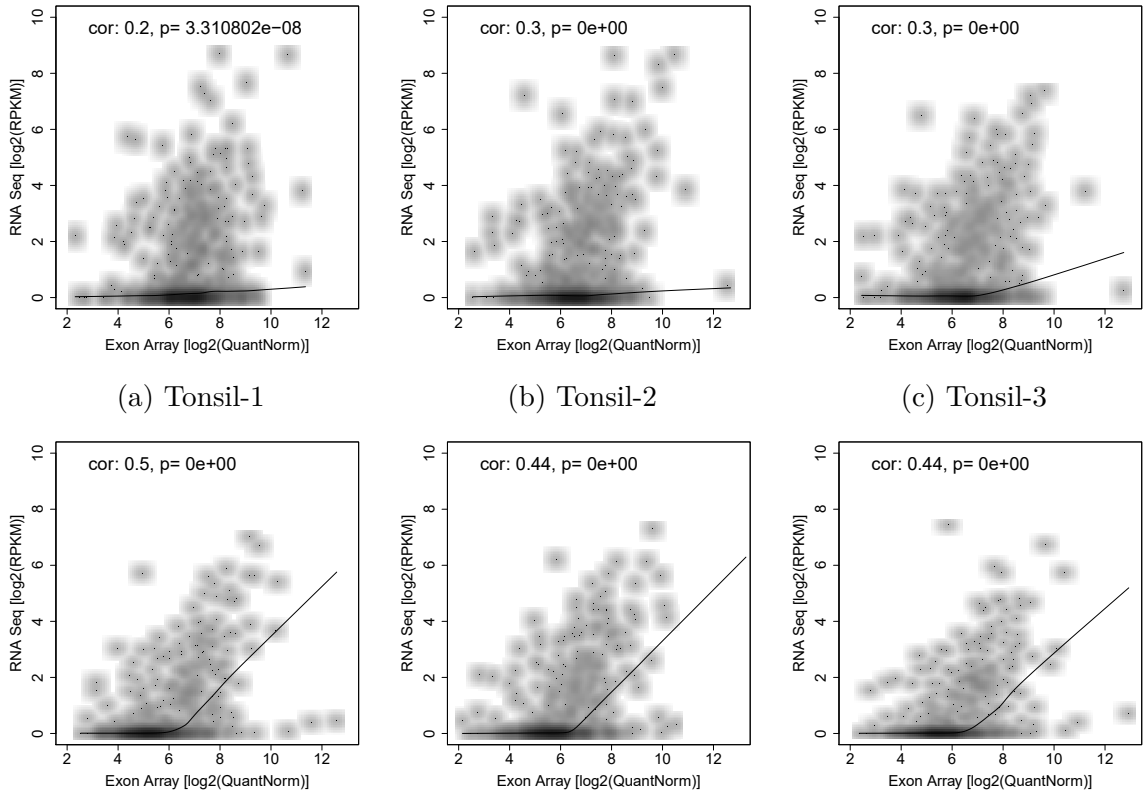
(d) CD44 - RNA sequencing

Figure 5.5: **CD44 and DRAM2 expression on probe level.** Probes denoted on the x-axes are sorted according to genomic positions. Samples are color-coded according to groups, an additional, thicker line denotes the group mean.

DRAM2 is either similarly expressed in all samples based on the sequencing data, or shows variation in individual samples, according to exon array data, but on a gene-wide basis, rather than for individual exons.

5.3.3 Gene Level

For comparison on the level of genes, we computed expression values as described in Section 5.2.1. Each of the two expression sets is then analyzed for differential gene



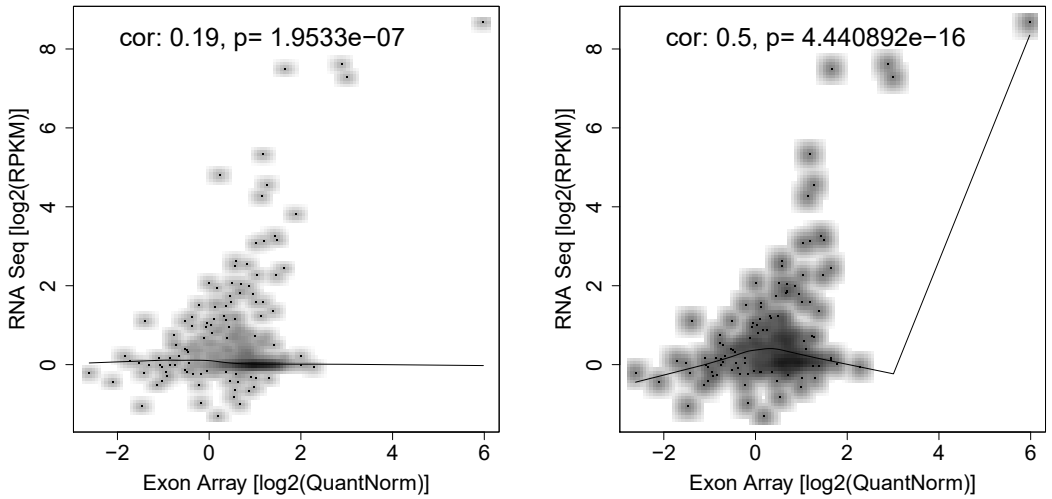
(a) Tonsil-1 (b) Tonsil-2 (c) Tonsil-3

(d) DLBCL-4 (e) DLBCL-5 (f) DLBCL-6

Figure 5.6: **Probe set level scatterplot of sample correlation.** The corresponding expression values from both technologies are compared per sample. Only probe sets uniquely mapped to exons are included. A lowess line is included in every plot for a better visualization of comparability.

expression according to the analysis methods suitable for the respective technology. We thus obtained a set of differentially expressed genes for the exon array data as well as the sequencing data, which we correlated by fold change. To assess the concordance in differential gene expression prediction, we intersected the two result sets. Note that we did not apply pre-filtering on the data, as the method used for differential gene expression on sequencing data (DESeq2) requires all unprocessed counts.

Based on the probe level results, where most of the probes matched to CCDS-exonic regions, but a considerable amount also matched to other regions inside a gene, we prepared a second gene level data set for the sequencing data, where we included all reads mapping to the whole gene region instead of the exonic parts only. With this, we wanted to assess whether the 'extended' whole-gene data set represents exon array data better. Details on the differential expression results from these two sequencing data sets are shown in Table 5.4. Differences in the number of up- and down-regulated genes (1,381 vs. 7,525 (up) and 847 vs. 6,296 (down)) for CCDS exon and whole gene respectively, as well as a changes in proportion (1.6 (exon) vs. 1.2 (whole gene) times



(a) fold change correlation (b) fold change correlation (high expr)

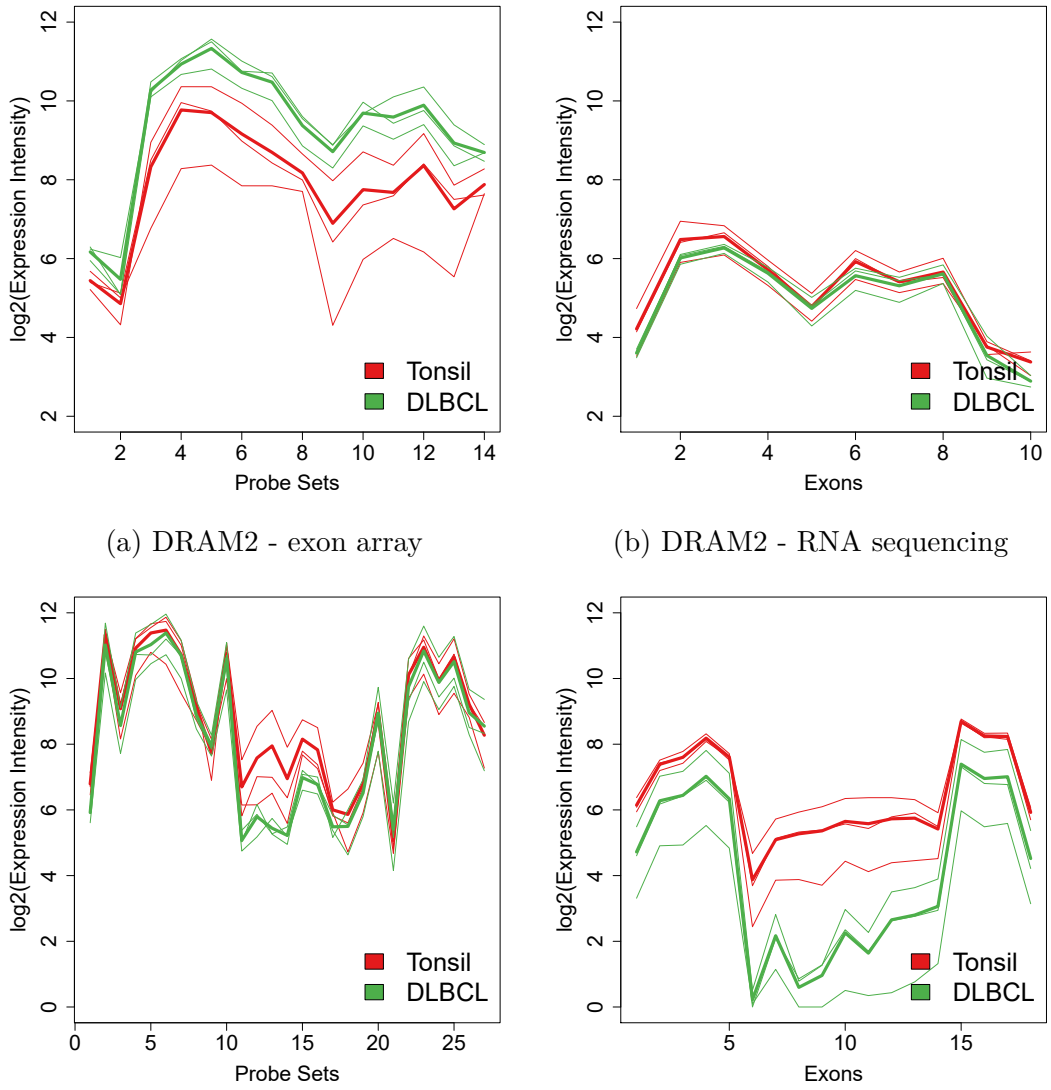
Figure 5.7: **Probe set level fold change correlation** without (a) and with (b) low expression filtering based on RNA sequencing data.

	DESeq2 results for	
	exon based data set	whole gene data set
nonzero total read count	26408	54230
LFC > 0 (up)	1381, 5.2%	7525, 14%
LFC < 0 (down)	847, 3.2%	6296, 12%
outliers	134, 0.51%	619, 1.1%
low counts	11998, 45%	8383, 15%
p-values (fisher)	0.01243525	2.007481e-17

Table 5.4: **Results for DESeq2.** Two different gene level data sets are used for the determination of differential gene expression based on RNA sequencing data. One count data set, generated with HTSeq, is based on the exons of a gene (left column), while the second, derived from the use of coverageBed, includes all reads mapping to intronic regions as well (right column). Additionally, the fisher test p-value for the significance of the intersect with the exon array gene level results on differentially expressed genes is displayed.

more upregulated) within the data sets can be observed. Exon array data analysis for differential gene expression using limma resulted in 3,221 significantly deregulated genes (Benjamini-Hochberg corrected p-value ≤ 0.05 and fold change ≥ 2). of which 1,230 were down and 1,991 were up-regulated in DLBCL.

Comparison of the two differential gene expression results based on sequencing data to the genes differentially expressed in exon array data shows higher comparability for



(a) DRAM2 - exon array

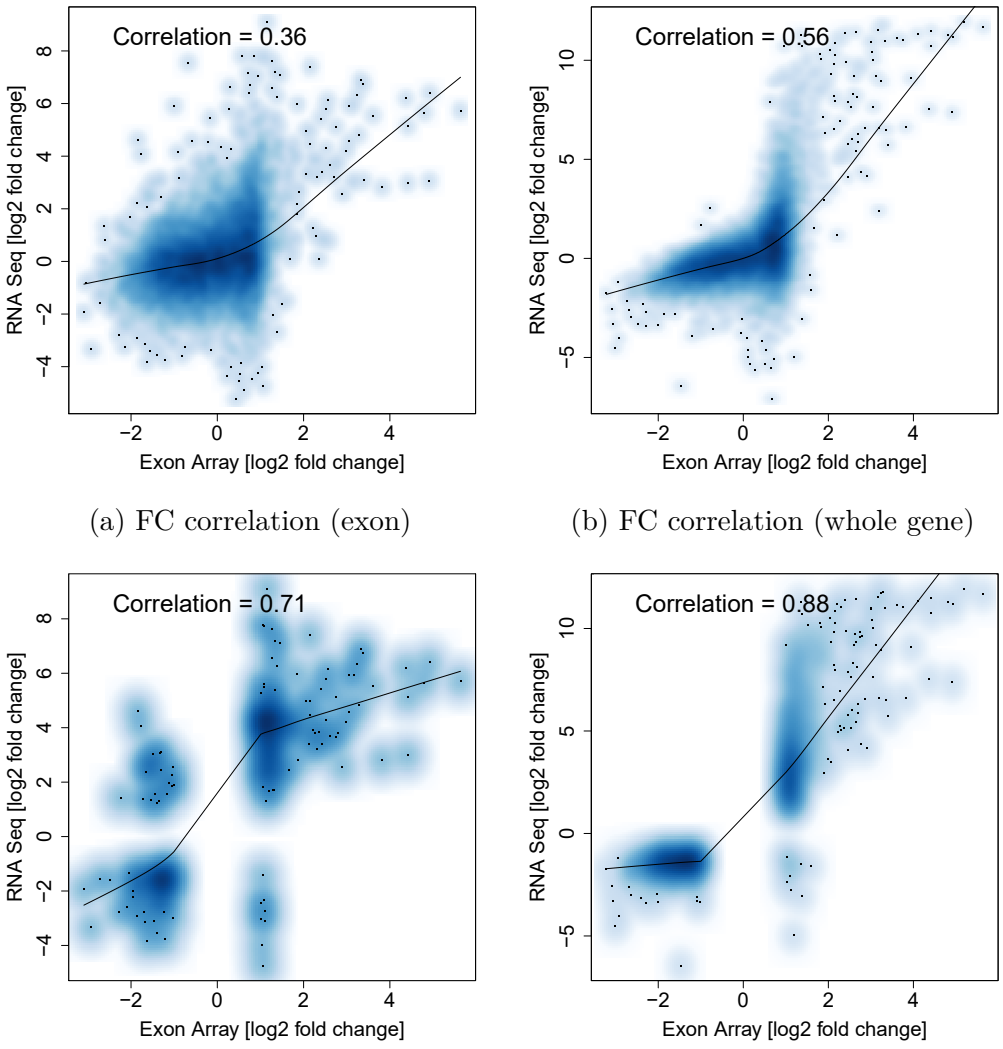
(b) DRAM2 - RNA sequencing

(c) CD44 - exon array

(d) CD44 - RNA sequencing

Figure 5.8: **CD44 and DRAM2 expression on probe set level.** Probe sets and exons denoted on the x-axes are sorted according to genomic positions. Samples are color-coded according to groups, an additional, thicker line denotes the group mean. For CD44 differential splicing is visible for both technologies. The difference between group means is greater in the middle region of the gene, which corresponds to the spliced region.

the whole-gene data set (see Figure 5.9). Concordance in result sets and correlation of fold changes improves for the sequencing set covering the whole gene region. For the set of all genes, correlation raises from $r = 0.36$ (Subfigure 5.9a) to $r = 0.56$ (Subfigure 5.9b). For the significantly differentially expressed genes only, correlation increases from $r = 0.71$ (Subfigure 5.9c) to $r = 0.88$ (Subfigure 5.9d). All correlations were highly



(a) FC correlation (exon)

(b) FC correlation (whole gene)

(c) FC correlation of DE genes (exon)

(d) FC correlation of DE genes (whole gene)

Figure 5.9: **Gene level comparison of fold changes.** For both gene level data sets, the exon-based (left picture column) as well as the whole gene-based (right picture column), fold changes between groups within one technology are computed and compared to the second technology.

significant. More importantly, the gene-wide count data set reduces the number of fold change direction switches for the significant genes as shown by the reduction of data points in the upper left and lower right quadrant in Figure 5.9d compared to Figure 5.9c.

The number of overlapping significantly expressed genes is visualized in Figure 5.10.Note that we included all genes (ensembl gene ids) found in both gene expression sets to be compared, i.e. sequencing (14,276 (exon-based) and 45,228 (whole-gene based)) and

(a) exon array vs seq(exon-based) $p = 0.01$

(b) exon array vs seq(whole gene-based) $p = 2.007481e - 17$

Figure 5.10: **Gene level comparison of differentially expressed genes.** For both sequencing gene level data sets differentially expressed genes (adjusted p-value < 0.05 and $|FC| > 2$) are assessed and compared to differentially expressed genes from the exon array data. Both overlaps are significant. Note that in each comparison, only genes present in both data sets compared are included.

exon array (14,871). Thus, as the initial sequencing data sets differ in terms of the genes they cover (for example due to low counts), also the set of genes compared in Figure 5.10a and Figure 5.10b differ. While this might bias results, we feel that it gives a clearer picture on the overall differences than working on a reduced set of genes which can be found in all three data sets. The overlap in differentially expressed genes between exon array based results and differentially expressed genes from data retrieved on the exonic parts of the gene only, shows a much higher p-value (fisher exact test) with $p = 0.012$ compared to the overlap based on the whole gene expression set ($p = 2.007481e - 17$).

5.3.4 Differential Splicing Level

Both of the methods applied in this step, ARH and SplicingCompass, make predictions on gene level, i.e. they rank genes based on the probability to have undergone a differential splicing event. Even though the third method, DEXSeq, operates on exon level, we have no means to compare exon level results to the other two methods. Thus, DEXSeq results are reduced to gene level predictions for comparison. Irrespective of the p-values computed for every gene, we determined fold changes on exon level for all methods which we used for filtering. Therefore, those genes are classified to experience differen-

tial splicing that exhibit at least one exon level fold change of 1.5 and a method-specific $p - value < 0.05$.

To assess the influence of Benjamini-Hochberg multiple testing correction on the concordance between methods, we used two result sets for comparison. We use a permutation test to compute p-values for ARH. As of the small sample size, the minimal p-value is limited. Thus, a multiple testing correction would classify all genes as not differentially spliced for pure technical reasons. To enable a fair comparison, we included all the methods applied without multiple testing correction and, in a second round, we compared SplicingCompass and DEXSeq with corrected p-values.

Applying DS detection methods on the data prepared accordingly (see also Section 5.2.1), led to a gene level output as shown in Table 5.5. ARH leads to the highest number of genes predicted to be differentially spliced, while SplicingCompass gives the most conservative prediction. The concordance based on numbers of overlap and corresponding p-values is shown in Table 5.8 for the unadjusted, and Table 5.6 for the adjusted case and is discussed in more detail in the following.

			Number of DS genes	
Method	Technology	Genome.version	p.value	p.adjusted
ARH	Exon Array	HG19	2508	-
SpComp	Exon Array	HG19	1805	246
DEXSeq	NGS	HG19	3316	912
ARH	NGS	HG19	3909	-
SpComp	NGS	HG19	762	56
DEXSeq	NGS	HG38	3759	1023
ARH	NGS	HG38	3819	-
SpComp	NGS	HG38	723	58

Table 5.5: Results for differential splicing detection. The number of genes with indication for DS by p-values and multiple testing corrected p-values (p=0.05) for all methods, technologies and human genome versions.

Comparing the **overlap between technologies** on the same genome version without adjustment yields six result sets. Two of them, namely the intersection between ARH (array(A)-HG19) and ARH (sequencing (S)-HG19) and SpComp (A-HG19) and ARH (S-HG19) are significant. None of the two adjusted result sets that are available for quantifying the comparability is significant. When considering the newer genome version HG38 instead of HG19 for sequencing data and the results without adjustment, four of the six intersections are significant. ARH (S-HG38) and SpComp (S-HG38) exhibits a significant overlap with results of ARH (A-HG19) and SpComp (A-HG19).

Note that DEXSeq displays no significant intersection with both array based results, yet, exon array results also lack a set based on the same method.

SpComp (S-HG38) shows a significant overlap with SpComp (A-HG19) while the intersection with ARH (A-HG19) is close to significance ($p = 5.465e-02$). However, none of the two adjusted result sets are significant.

The **overlap within technologies** is high in the case of exon arrays, i.e. SpComp (A-HG19) and ARH (A-HG19). Note that this is the only intra array comparison scenario. For the three sequencing cases, a significant overlap is observed for the comparison of SpComp (S-HG19) with ARH (S-HG19) as well as with DEXSeq (S-HG19). The same result is obtained when comparing sequencing data based on HG38, SpComp (S-HG38) and ARH (S-HG38) as well as SpComp (S-HG38) and DEXSeq (S-HG38) have a significant intersection.

When comparing **genome versions**, nine intersection can be used on basis of sequencing data. Besides the comparison of the identical methods, which all had an highly significant p-value, both comparisons between ARH and SplicingCompass were significant. For ARH (S-HG38) versus SpComp (S-HG19) as well as for ARH (S-HG19) versus SpComp (S-HG38) the p-value was highly significant. Note that the only two significant comparisons in the adjusted case, were the ones between identical methods with differing genome versions, i.e. the comparisons expected to be highly similar. Figure 5.11 displays high comparability between genome versions, given the rather stable numbers in the intersections between HG19 (Figure 5.11a) and HG38 (Figure 5.11b).

Looking at the **overlap between methods** on the three different comparison levels A-HG19, S-HG19 and S-HG38, i.e. for fixed technologies and genome versions, yields very consistent results. Each of the three levels shows a significant overlap between ARH and SplicingCompass and if present, SplicingCompass and DexSeq. DEXSeq and ARH never overlapped significantly.

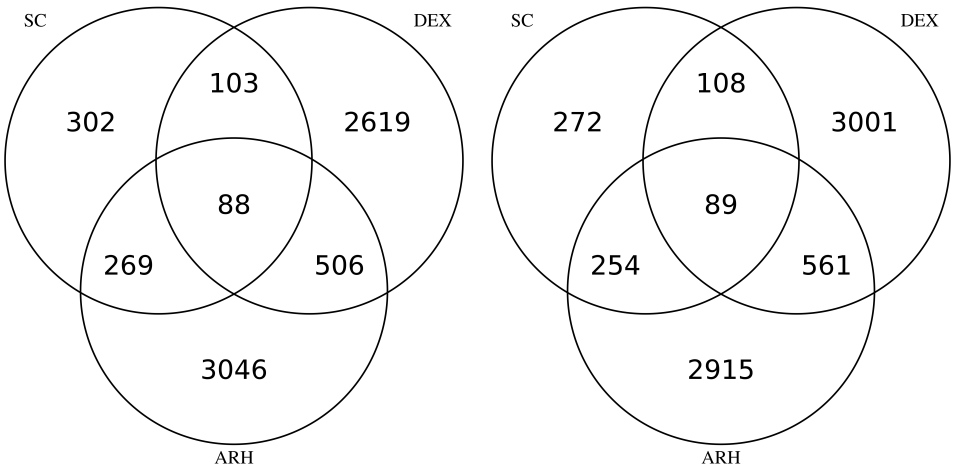
The **overlap within methods**, i.e. the comparison of the same method between different technologies and genome versions, provides three intersections for ARH and SplicingCompass and one for DEXSeq. For ARH and DEXSeq all intersections are significant. SplicingCompass results only in significant overlaps for two out of three comparisons, the array-sequencing approach based on HG19 is not significant in overlap. For the corrected scenario both SplicingCompass and DexSeq are significant in their overlaps. An overview on the intersections between the three result sets for ARH is given in Figure 5.12a and for SplicingCompass in Figure 5.12b, respectively.

We also assessed the overlap between all methods, genome versions and technologies for both the adjusted and unadjusted scenario. The unadjusted set contained 7 genes shown in Table 5.7. For the adjusted results, only one gene, COX11 cytochrome c oxidase copper chaperone, was identified in all sets.

5.3.5 The Impact of Filtering

The different levels of data comparison identify several characteristics of data based on the two technologies. We thus sought to determine whether filtering can improve the concordance for results in differential splicing between technologies.

To this end, we first compared data on the probe level. Based on the expression distributions shown in Figure 5.4, we hypothesized that removal of the low expressed



(a) HG19 (b) HG38

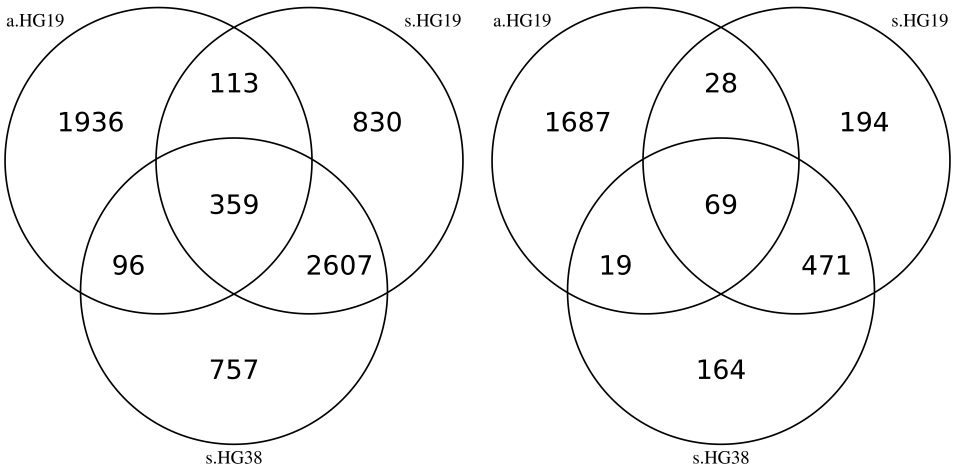
Figure 5.11: **The impact of the genome version.** Result overlap between methods for DS detection based on RNA sequencing data annotated with HG19 and HG38 respectively.

	SpComp (A-HG19)	DEXSeq (S-HG19)	SpComp (S-HG19)	DEXSeq (S-HG38)	SpComp (S-HG38)
SpComp (A-HG19)	0	2.028e-01	1.000e+00	2.161e-01	1.000e+00
DEXSeq (S-HG19)	15	0	1.537e-01	0.000e+00	7.771e-01
SpComp (S-HG19)	1	6	0	7.661e-02	2.835e-46
DEXSeq (S-HG38)	16	749	7	0	5.577e-02
SpComp (S-HG38)	1	4	25	8	0

Table 5.6: **Size of result overlaps and corresponding p-values with p-value correction.** The lower triangle shows the number of overlapping genes predicted as DS based on the corresponding methods and genome versions. The upper triangle displays the respective p-values. 'A' (array) and 'S' (sequencing) denote the technology used.

probes based on RNA sequencing data might increase expression correlation. We thus removed the lower half of the probes based on the sequencing expression values in both data sets. This procedure slightly worsened sample-wise expression correlation, as did the removal of the lower expressed half of probes based on exon array data.

Based on this finding, we assessed the general relationship of expression correlation and comparability of differential expression. More precisely, we applied filtering of entities showing low expression (1) based on exon array data and (2) based on RNA sequencing data on the exon as well as on gene level and compared correlation of expression as well as correlation of fold change to assess whether this procedure improves result concordance.



(a) ARH (b) Splicing Compass

Figure 5.12: **Differential splicing result comparability of each method throughout technologies and genome versions.** Genes classified as differentially spliced according to ARH and SPCOMP are intersected for each of the methods in all applied scenarios.

SYMBOL	DESC	ENSEMBLID
DTNB	dystrobrevin beta	ENSG00000138101
AKAP9	A-kinase anchoring protein 9	ENSG00000127914
IGF2BP3	insulin like growth factor 2 mRNA binding protein 3	ENSG00000136231
WHSC1L1	Wolf-Hirschhorn syndrome candidate 1-like 1	ENSG00000147548
RBM26	RNA binding motif protein 26	ENSG00000139746
XYLT2	xylosyltransferase II	ENSG00000015532
COX11	COX11 cytochrome c oxidase copper chaperone	ENSG00000166260

Table 5.7: **Genes predicted as differentially spliced in all technologies, methods and genome versions.** The intersect is derived from unadjusted results, for adjusted p-values, only one gene, COX11, is contained in every result set.

For the probe set, i.e. exon level, results based on expression correlation showed a group-dependent behavior. Sample-by-sample correlation improved for Tonsils, while DLBCL showed a slightly decreased correlation. Fold change correlation improved after filtering of low expression, when based on array data, to $r = 0.29$ and based on sequencing data to $r = 0.5$.

For gene level data, expression correlation decreased for all samples. When filtering based on array genes, correlation of fold changes achieved $r = 0.43$, and $r = 0.31$ in the case of filtering based on sequencing data. Note that the unfiltered baseline is $r = 0.36$.

For detection of differential splicing, filtering on gene basis instead of exon basis is advisable. A gene has to be fully represented or discarded from analysis, to avoid genes with only partial representation of their exons. Thus, we filtered DS results by keeping only genes showing high expression in exon arrays and genes showing high expression in sequencing data. The amount and significance of result overlap is compared to the unfiltered results. For filtering based on sequencing data, results deteriorated. Three overlaps formerly significant now had a p-value > 0.05. Additionally, the p-values of those comparisons still significant increased. When filtering based on array data, p-values also increased for most of the comparisons, but four more overlaps were significant compared to the unfiltered scenario (see Table 5.9).

5.3.6 Factors Impacting DS Comparability

Several methods, technologies and genome versions lead to a multitude of scenarios compared. To quantify the variables impacting the comparability of results on differential splicing level most, we fitted a generalized linear model to the three factors enumerated. Method, technology and genome version were encoded, with respect to their concordance (x in the case of consensus, y otherwise) and set into relation to the binary outcome, i.e. 1 if the overlap was significant and 0 otherwise.

The most impacting variable was the method (p=0.06), i.e. whether the two result sets intersected were produced by the same method or not. The second important factor impacting significance of overlaps was the fact that the two sets compared were derived from the same technology ($p = 0.07$). The genome version showed no significant impact on the comparability of results ($p = 0.9$).

5.3.7 Application to a Glioblastoma Multiforme Data Set

We selected a second, publicly available data set, for which exon array data and RNA sequencing data were available for a set of samples to validate our framework. This data set is based on four glioblastoma multiforme samples (GBM), the most aggressive cancer originating in the brain, and four organ specific control samples derived from TCGA [Tomczak et al., 2015]. As opposed to the DLBCL data set, the RNA sequencing GBM data is based on paired-end sequencing. A list of the samples used is given in Table 6.14 (Appendix).

We applied our multi-level framework to the GBM data set and put results side-by-side to results obtained from DLBCL data (see Table 5.10). Concordance of RNA sequencing and exon array analyses is improved on every level. All correlations presented in the following are highly significant.

- For probe level and probe set level, sample-wise Pearson expression correlation is higher in the control group of GBM, while for DLBCL, expression correlation for cancer samples exceeds the control group.
- Filtering of low expression data on probe set level leads to consistent results for both data sets, fold change correlation improves even more for filtering based on sequencing data (DLBCL: $r = 0.5$, GBM: $r = 0.71$) compared to array data (DLBCL: $r = 0.29$, GBM: $r = 0.70$).
- For the gene level, result comparability is based on exon-derived gene counts. A higher concordance in the GBM data set is observed for fold change correlation as well as for the overlap in differentially expressed genes.
- The overlap in predictions of differentially spliced genes based on array and RNA sequencing data is in line with results of the previous levels of the GBM data set. All overlaps between the two methods applicable for both technologies are significant. This is contrasted by the DLBCL results for SplicingCompass on RNA sequencing data, which show no significant overlap with array-based results.
- For the multiple testing corrected differential splicing results, the intersection between differentially spliced genes based on SplicingCompass (array) and Splicing Compass (seq) was significant (p=1.027e-03). Note that we again did not apply multiple testing correction on results based on ARH for the same reasons as for the DLBCL data set.

5.4 Discussion

As RNA sequencing is more and more replacing exon arrays in the detection of differential splicing but many previous and important results are based on exon arrays, knowledge about the comparability of the two technologies is crucial. First and foremost it is of high interest, to understand whether the newer technology can be used synonym to the older, and if not, what the restrictions are. This is also a relevant question with regard to, second, comparisons of RNA sequencing results to published exon array results from literature and the corresponding public databases.

In this section, we presented a multi-level framework for the comparison of differential splicing results based on exon array data and RNA sequencing data derived from the same biological samples. For a deep and detailed understanding of comparability and the related pitfalls, we implemented comparisons on various levels of analysis: From the basic probe level, to probe set expression, i.e. exon level comparison, gene level expression as well as differential splicing.

Comparison on different levels was key in understanding the full nature of the data, and provided valuable filtering approaches ameliorating result comparability. Even though lower levels of our framework reveal only moderate correlation in sample wise expression as well as in fold change, a relevant concordance in prediction of differential splicing is observed. More importantly, our results give room for the hypothesis

that the application of the same DS detection method has a similar impact on result comparability of DS as equality of underlying technologies does.

5.4.1 Results on Different Levels.

In the following, we discuss results for the different levels.

Probe Level. On the most basic level of comparison, the *probe level*, concordance of results can be assessed on the most controlled data, as we explicitly computed only those counts for the sequencing set corresponding to the exact probe sequence used on the microarray. Thus, we expect to see only differences induced by technology and sample-taking procedures as none of the further aggregation steps are applied. Note that all further comparison levels are based on values which are acquired differently, i.e. based on the standard protocol for the respective technology and analysis method. This means that expression values for exon arrays are always an aggregate of probe values while sequencing data exactly measures the expressed transcripts for the whole entity under consideration. For these reasons, we expected expression correlation to be the most concordant on this level.

However, results reveal a rather moderate overall correlation, which can be traced back to the distributions of expression values for the respective technologies. The differing scales, partly responsible for low correlation, are best shown in low expression regions (see Figure 5.4). Data points which spread over a large scale in count data have a much denser expression range for exon array data, which, together with additional noise induced, for instance, by cross-hybridization and different probe affinities in exon arrays, is counteracting a high correlation. As shown in Figure 5.5, exon array data displays a high variation (between probes) on this level, which is not found in RNA sequencing data and this difference impacts correlation substantially.

An additional source of error is induced by a characteristic of RNA sequencing data. Expression of transcripts in low expression regions are less reliable, as it can not be doubtlessly determined whether the expression observed is due to biological reality or a coverage bias induced by potentially overrepresented transcripts.

Even though designed on HG16, most of the core probes (about 90%) of the exon array uniquely map into HG19. Of these, a relevant amount, about 7%, do not map to annotated CCDS exons. The impact of these probes is further discussed for the gene level.

Probe Set Level. The *probe set level* exhibits a group specific expression correlation change: tonsil correlation increases, while DLBCL correlation slightly decreases. Even though this level incorporates the lowest amount of data points, expression correlation ranges over levels are consistent and thus proof reliability. The most interesting observation is the low correlation of fold changes. While on the probe level underlying genomic regions interrogated are unified and on gene level a much higher amount of measuring points is aggregated to one, the probe set level struggles from injective mappings and

thus aggregation of data points with a high underlying variability. Also, the different scales of the two technologies contribute to the result observed, as shown by improvement through low expression filtering. The relative low number of data points might be due to the exon array probe design. Probes are known be ambiguous sometimes [Liu et al., 2010], which impedes the unique mapping of probe sets to exons.

Visualization of CD44 and DRAM2 on the probe as well as the probe set level provides two valuable insights (see Figure 5.5 and 5.8). First, the high variance observed for probe level exon array data explains a relevant part of the low expression correlation observed between corresponding samples on the probe level. Second, the probe set level shows the necessity for aggregation of expression values, as this level displays a much higher comparability in expression pattern and variance betwenn technologies. Also, DS for CD44 in both technologies is visually observable only for the probe set level, as here the group means differ in the middle region of the gene. Note, that this difference is much more pronounced for the sequencing data, as Figure 5.8d displays a higher relative difference in the spliced regions compared to Figure 5.5d. This might be due to the fact that more information is included in RNA sequencing data on the exon level than on probe level, as the probe level exclusively relies on regions covered by uniquely mapping probes.

Gene Level. The use of two different data sets on the *gene level* demonstrates the impact of the intronic and non CCDS exonic probes as well as the impact of coverage. Across all comparisons, the data set built from count data in the whole gene regions shows higher concordance with the gene level exon array data. Both the overall correlation of FCs as well as the correlation of the significant genes only increase with the whole gene data set. The intersection of differentially expressed genes on basis of the whole-gene data shows a much lower significant p-value. Nevertheless, both overlaps are significant. A further aspect pointing to a better representation of the exon array by the whole-gene data set is the observation of highly reduced fold change contradictions. While Figure 5.9c displays a relevant number of fold changes with differing signs, these are reduced to a minimum in Figure 5.9d.

Besides additional probes, two other sources might influence the outcome. Note, (1) that the two sequencing data sets are necessarily generated by different tools. Also, (2) the differing number of genes included due to low coverage can influence correlation. Compared to the probe set level, correlation of fold changes is more stable, which hints to the general rationale of transcription microarrays: more measuring points are more reliable and represent entities better.

Level of Differential Splicing. For the comparison on the *level of differential splicing* we relied on the result intersections between methods and technologies. While on the uncorrected results significance in overlaps can be observed for several comparisons, the multiple testing corrected result set shows no relevant overlap between results of different technologies or different methods. We thus focus on the comparison of the uncorrected

result sets, knowing that they contain a relevant number of false positives. Nevertheless, the tendency they show is valuable and cohesive.

The most interesting comparisons are the overlaps between technologies based on methods applied in both. For HG19, two out of four sets are significantly overlapping. Interestingly, the two non-significant intersects are the ones containing SplicingCompass results, which is the only method exploiting splice junctions in this approach. For HG38, three out of four comparisons have a significant intersection. The additional one is based on the same method, SplicingCompass. An important evidence for the impact of using the same method is the observation that DEXSeq has no significant overlaps with array based data. Nevertheless, technology does have a great impact, as, for instance, significant overlaps between DEXSeq and SplicingCompass, when based on the same technology, show.

5.4.2 The Impact of Splice Junctions.

The methods applied for detection of differential splicing differ significantly in their approaches. While ARH uses an information theoretical approach, SplicingCompass attempts to solve the task by using analytical geometry. DEXSeq uses a generalized linear model to test for differences in variance on basis of a negative binomial distribution. Another major difference is the use of splice junctions. For exon arrays, no such probes are available. DEXSeq does not make use of the information directly, only by adding counts derived from spliced reads to the different exons they belong to. ARH is in theory able to include such information for the sequencing scenario, though, for comparability reasons, we decided to not use the information in our approach. Thus, SplicingCompass applied on sequencing data is the only method using exon junction information. This is reflected in the insignificant overlaps between SplicingCompass on sequencing data and the array based results, as the use of splice junctions reduces the number of genes classified as differentially spliced. While based on exon array data the number of differentially spliced genes is rather high, it reduces significantly for the sequencing data. Numbers are lower as all the predictions of both other methods, and they further decrease to a minimum compared to DEXSeq, when multiple testing correction is applied (see Table 5.5). Thus, results on DLBCL data suggest that splice junctions have a profound impact on the comparability of results. For the GBM data set, both SplicingCompass based intersections are significant. However, they display the highest p-value compared to the remaining overlaps.

5.4.3 The Homogeneity of Tumor Data.

Tumor heterogeneity is a wide research field, not only, but also manifesting on the level of gene expression [Marusyk and Polyak, 2010]. Intra- as well as inter-tumor heterogeneity is known, giving rise to different treatment options and necessities. As of this facts, we expected a higher concordance in the tonsil samples, which we do not see in our data. Interestingly, DLBCL samples correlate better throughout levels, even though not the same RNA extract is used for both technologies. Nevertheless, the biological material

originates from the same biological sample. For the GBM data set, the opposite holds, expression correlation is better for the control samples.

5.4.4 HG38 versus HG19 - the Influence of the Genome Version.

The latest genome version unofficially supported by Affymetrix at the time of analysis is HG19. Thus, our main comparison is based on this version. Nevertheless, we want to include current development and thus decided to additionally run analyses for sequencing data based on HG38. The main difference between HG19 and HG38 is the attempt to include information on more individuals in the reference genome. This is mainly reflected in various alternative sequences additionally provided. However, additional improvements in HG38 make it a better assembly with less gaps, and a higher quality in terms of the N50 scaffold. As far most of the aligners are not designed to make use of the alternative sequences, including the splicing-optimized aligner we used, we only benefited from the general improvements of HG38. The application of a different aligner able to include alternative sequences would also have de-focused our aim to compare the two technologies as fair as possible.

For the inner-sequencing comparison, i.e. the intersection between differential splicing results based on different genome versions, the effect is clearly dictated by methods. While the overlap between identical methods is highly significant for all three inner-method comparisons, three out of six inter-method comparison show significant results.

ARH and SplicingCompass have relevant similarity in their results produced, irrespective of the genome version. SplicingCompass and DEXSeq inter-genome version overlaps are significant or close to significant, while DEXSeq and ARH do not share relevant predictions in common. The concrete numbers in overlap between the three methods are - as expected - similar, yet by far not identical, for the comparison between genomes (as shown in Figure 5.11).

Interestingly, the genome version makes a relevant difference for the technology comparisons. While all overlaps between DEXSeq and ARH lack a relevant size, p-values for the remaining four sets (ARH vs. SplicingCompass) decrease for Seq-HG38 compared to Seq-HG19. In the case of SplicingCompass (ArrayHG19) versus SplicingCompass (SeqHG38) to the level of significance and for the intersection between ARHArray19 and SplicingCompassSeqHG30 to a level close to significance.

One explanation of this observation might lie in the design of the exon arrays. As observed in the more basic levels of our framework, most probe sets can not be bijectively mapped to Ensembl Exon IDs. Thus, all methods are applied on the probe set values of a gene. For most of the cases, this leads to an increase in the number of 'exons' per gene. Together with the fact that the exon array partly covers intronic regions, this might produced a more fine-grained resolution than by working on exon level only. While this could be a problem for purely statistical approaches, ARH is designed to perform independently of the number of exons. On the other hand, HG38 is an improvement in several aspects in comparison to HG19. Thus, we hypothesize that the more fine-grained exon level approach is better reflected by the more accurate genome version which in terms leads to better concordance. One evidence for this theory is the fact that ARH on

the exon array level forms significant overlaps with all three result sets from sequencing data based on HG38.

5.4.5 What Impacts Results?

In the following, we will discuss two major areas which impact result comparability. On the one hand, we have general factors like the methods applied or the underlying technology. Additionally, filtering impacts results and thus their concordance.

The Impact of Technology and Method. Obviously, the technology applied influences results in a major way. RNA sequencing and exon arrays produce expression values on different scales implying different levels of resolution. It is questionable whether, even in theory, all relations can be preserved. Differences might, especially in low expression regions, easily be overlaid by various effects such as cross-hybridization, background noise, probe affinities, amplification artifacts, and others. While in theory, this question might be best answered on the probe level, as the prerequisites for comparability are optimized in our probe level approach, variation in exon array data here is much higher compared to RNA sequencing data. This effect is impressively illustrated in Figure 5.5, while tendencies are the same in both technologies, RNA sequencing data is by far more stable and conclusive which obviously inhibits good correlation of results. Variance stabilization is induced by aggregation of measuring points. For the next-higher level, concordance in probe sets / exons is much higher as displayed in the example of Figure 5.8. This trend continues with higher aggregation, as FC correlation is even better on gene level. On the differential splicing level, both aggregation levels come into play, as exon level values are set into relation to the corresponding gene expression. The later additionally stabilizes expression values.

It is known, that methods for detection of differential splicing differ, at times vastly, in their predictions. Together with the bias induced by different technologies, most publications report rather low result concordance between RNA sequencing and exon array data when it comes to differential splicing [Raghavachari et al., 2012]. We thus wanted to assess, whether the application of the same methods to both kinds of data can significantly improve result concordance, i.e. whether the effect of differing technologies can be overcome by the choice of the adequate method.

According to our findings it is more important, if the same method is applied instead of the same technology. 3/8 (37,5%) of the significant overlaps are based on the same method while 1/12 (8,3%) of the non-significant intersects are based on the same method. For technologies, 5/12 (41,67%) of the non-significant intersects are based on the same technology, while 13/16 (81,25%) of the significant overlaps are based on the same technology. Thus, the effect of the same method is more than 4 times higher compared to the two-fold increase in technology impact. A model fit to results based on the potentially impacting factors confirm this finding, while both, technology and methods impact results, the effect of choosing the same method might be more important.

The Impact of Filtering. To enhance results and their concordance, several filtering approaches could be applied. Yet, 'over'-filtering has to be carefully avoided for several reasons. First and foremost, no gold standard is known, thus, no decision on 'correct' results can be taken. Second, extensive filtering could eliminate the advantages of the *high-throughput* approach or reduce coverage of genes drastically. Due to the latter, filtering should take place on gene level as opposed to probe or probe set/exon level, to avoid a sparse coverage of genes examined. As we see an application of our approach in the comparison of newly analyzed RNA sequencing data with older, published work, we claim an easy to apply and generic filtering a prerequisite. We believe that a sophisticated re-analysis is beyond the scope of researchers interested in comparing their results to pre-existing knowledge. Further, knowledge of the dataset based on the 'other' technology should not be necessary, i.e. filtering should be independent.

We thus applied filtering on the different levels to acquire knowledge on comparability improvement. Results showed no consistent improvement for expression correlation, but indicated improvements in the correlation of fold changes. We thus filtered DS results for lowly expressed genes based on array as well as on RNA sequencing data. While RNA sequencing based filtering decreased result overlaps, array based filtering lead to an increase in the number of significant overlaps.

5.4.6 Results for Different Data Sets

Comparison of results from our multi-level framework for the DLBCL and the GBM data sets showed an higher overall concordance for the GBM data set. We attribute this to the different sequencing mode, paired end versus single end, as well as the fact that RNA used for the DLBCL data set was extracted a substantial time apart for the two technologies. Despite the higher absolute concordance for the GBM data, the trend is similar for both data sets. Aggregation improves correlation for sample wise expression as well as for fold changes. Filtering of low expression values based on sequencing data improves fold change correlation more than array based filtering. Even though more significant overlaps on the level of DS are observed for the GBM data set, the p-value based ordering is mainly preserved between data sets.

5.5 Conclusion

RNA sequencing is more and more replacing exon arrays in the application of differential splicing detection. In this work, we aimed at understanding to what extent the two technologies are comparable with respect to this task. As current research lacks meaningful examination scenarios and classifies concordance as rather low [Raghavachari et al., 2012, Bradford et al., 2010], we tackled the problem by two major approaches. First, we developed a detailed multi-level framework examining and comparing RNA sequencing and exon array data in detail on every level. Second, we chose methods for differential splicing detection which are applicable to both data types, hypothesizing that a

great amount of divergence in DS prediction may result from the application of different methods.

Our multi-level comparison framework contributes substantially to the elucidation of technology-based data characteristics and comparability. It shows that high expression correlation on all levels is not necessarily a prerequisite for concordance in differential expression over levels. Even more, high expression correlation is rather unlikely on probe level, as array data is subject to high variation in all expression ranges. More stable values emerge by aggregation, which in turn improves comparability of differential expression events.

Filtering of low expression values can improve result concordance. Nevertheless, it poses the risk of loosing true positives and has thus to be carefully applied with the study aim in mind. Moreover, several methods for RNA sequencing data require unfiltered data (DESeq) or apply internal filtering themselves (SplicingCompass).

Result trends are similar for the GBM and the DLBCL data set, yet, result concordance is higher for the GBM data. We attribute this to the sequencing mode and the time points of RNA extraction. Paired end reads, used for sequencing of the GBM data (opposed to single end reads for the DLBCL data) may improve the overall quality of the data set, as they are, for instance, more likely to align to a reference. Moreover, differing timepoints for the extraction of RNA can influence result comparability.

The technology applied does have a slight impact on the comparability of DS results, nevertheless, the choice of the DS detection method might impact result concordance more. In our data, we observe several significant result overlaps. Even though for multiple testing corrected results no significant overlaps are achieved for the DLBCL data set, common tendencies clearly manifest on the uncorrected level.

Nevertheless, the two technologies can not be used synonymously, as a significant overlap is, in most cases, not based on an equal result set. Factors such as differing ranges of expression values, mostly in the lower expression levels, impact comparability. Moreover, the use of splice junctions, which is crucial for the final proof of an observed splicing event, is restricted to RNA sequencing data.

	ARH (A-HG19)	SpComp (A-HG19)	DEXSeq (S-HG19)	ARH (S-HG19)	SpComp (S-HG19)	DEXSeq (S-HG38)	ARH (S-HG38)	SpComp (S-HG38)
ARH (A-HG19)	0	0.000e+00	8.180e-01	3.339e-07	2.044e-01	7.534e-01	5.552e-09	5.465e-02
SpComp (A-HG19)	1638	0	8.257e-01	4.866e-11	1.108e-01	1.000e+00	1.585e-11	3.690e-02
DEXSeq (S-HG19)	425	322	0	4.970e-01	2.997e-02	0.000e+00	2.876e-01	6.041e-02
ARH (S-HG19)	472	306	594	0	2.721e-76	1.370e-01	0.000e+00	8.129e-58
SpComp (S-HG19)	127	97	191	357	0	3.087e-03	4.440e-75	0.000e+00
DEXSeq (S-HG38)	473	355	2850	647	209	0	5.467e-01	9.895e-03
ARH (S-HG38)	455	299	575	2966	348	650	0	1.241e-80
SpComp (S-HG38)	116	88	179	301	540	197	343	0

Table 5.8: **Size of result overlaps and corresponding p-values without p-value correction.** The lower triangle shows the number of overlapping genes predicted as DS based on the corresponding methods and genome versions. The upper triangle displays the respective p-values. 'A' (array) and 'S' (sequencing) denote the technology used.

	ARH (A-HG19)	SpComp (A-HG19)	DEXSeq (S-HG19)	ARH (S-HG19)	SpComp (S-HG19)	DEXSeq (S-HG38)	ARH (S-HG38)	SpComp (S-HG38)
ARH (A-HG19)	0	0.000e+00	1.907e-01	5.800e-03	4.864e-01	2.406e-01	1.069e-03	4.254e-02
SpComp (A-HG19)	427	0	1.754e-01	2.843e-02	4.332e-01	2.509e-01	5.561e-03	4.781e-02
DEXSeq (S-HG19)	193	149	0	1.441e-01	9.509e-03	0.000e+00	2.247e-02	1.515e-02
ARH (S-HG19)	90	69	184	0	1.489e-26	1.293e-02	1.137e-274	3.486e-19
SpComp (S-HG19)	39	28	99	107	0	9.665e-03	5.117e-28	8.875e-202
DEXSeq (S-HG38)	211	162	1009	188	103	0	2.262e-02	6.723e-03
ARH (S-HG38)	88	66	180	474	110	197	0	6.443e-27
SpComp (S-HG38)	31	22	95	91	187	101	104	0

Table 5.9: **Size of result overlaps and corresponding p-values without p-value correction for low expression filtering based on array data.** The lower triangle shows the number of overlapping genes predicted as DS based on the corresponding methods and genome versions. The upper triangle displays the respective p-values. 'A' (array) and 'S' (sequencing) denote the technology used.

	Data sets	
	DLBCL	**GBM**
Probe level		
sample wise correlation (cancer)	{0.47, 0.48, 0.54}	{0.49, 0.53, 0.57, 0.59}
sample wise correlation (control)	{0.11, 0.25, 0.29}	{0.57, 0.58, 0.58, 0.59}
Probe set level		
sample wise correlation (cancer)	{0.44, 0.44, 0.5}	{0.66, 0.68, 0.7, 0.71}
sample wise correlation (control)	{0.2, 0.3, 0.3}	{0.7, 0.72, 0.72, 0.73}
fold change correlation	0.19	0.64
fold change correlation (seq-filter)	0.5	0.71
fold change correlation (array-filter)	0.29	0.7
Gene level		
fold change correlation	0.36	0.77
fold change correlation (DE genes)	0.71	0.93
p-value for intersection of DE genes	0.01	0.006
DS level (p-value overlap)		
arhA-spcompA	0.000e+00	0.000e+00
spcompS - arhS	2.721e-76	2.016e-189
arhA - arhS	3.339e-07	5.784e-11
arhA - spcompS	2.044e-01	1.709e-04
spcompA - arhS	4.866e-11	2.498e-14
spcompS - spcompA	1.108e-01	2.786e-04

Table 5.10: **Comparison of the multi level results for the diffuse large B-cell lymphoma (DLBCL) and the glioblastoma multiforme (GBM) data set.** The DLBCL data set contains three control and three cancer samples while the GBM data set comprises four samples for each group. Samples from the DLBCL data set are based on single end sequencing while samples of the GBC data set are based on paired-end sequencing. Filters are applied either on array (array-filter) or on RNA sequencing (seq-filter) basis; the lower half of expression values based on the indicated technology is removed in both data sets. For DS comparison, the p-value of the result overlap based on methods (indicated in lower case letters) and technologies (A for array, S for sequencing) is displayed for both data sets.

6 Summary and Outlook

The main aim of this work was to advance the elucidation of the role of differential splicing in cancer. A detailed knowledge about (A) the occurring differential splicing events and (B) the main regulators provoking aberrant splicing profoundly improves the understanding of the underlying pathological mechanisms and can in consequence, improve therapeutic approaches.

In this work, we compared different methods and technologies for detecting differential splicing events and proposed a new method for the identification of aberrant splicing regulators. Assessing the performance of different DS detection methods with respect to several global data parameters, such as the number of samples per group, enables an informed method choice based on the data. We developed a novel method for the detection of potentially causal regulators amongst splicing factors based on transcriptomic data, while predicting candidates with alterations on potentially several influencial levels. Moreover, we proposed a multi-level framework enabling a comprehensive assessment of the concordance of exon array and RNA sequencing based detection of differential splicing, thus providing a profound evaluation of their comparability.

Chapter 3 described the comparison of nine methods for differential splicing detection based on exon arrays. A variety of methods exist for this task, but result overlap is often low. Furthermore, no comparative assessment addressing the susceptibility of the different methods to global data parameters has been conducted so far. Hence, we evaluated various methods available for differential splicing detection with respect to different data parameters. To this end, we generated artificial data sets with an altered number of samples per group, varying exon numbers per gene, different expression intensity as well as a varying percentage of samples differentially spliced per group. Additionally, we applied all methods to two published data sets for which experimental validation of predicted splicing events was available. We assessed accuracy, sensitivity and specificity for each method and data set. Furthermore, we determined the influence of the different global data parameters represented by the artificial data sets and provided a detailed discussion on the performance differences of the distinct methods with respect to the underlying algorithms.

Our results obtained for artificial and experimental data confirm and complement each other and are thus providing a global view on the performance of the different methods as well as their susceptibility to the data parameters implemented. According to our results, all methods are influenced by expression intensity as well as by the percentage of differentially spliced samples per group. Parameters such as the number of exons per gene allows for a distinction between methods responsive and unresponsive to this parameter in terms of accuracy. The same holds for the number of samples per group. If the sample number is low or unbalanced we recommend SplicingCompass. For

independence on the number of exons in a gene, ARH, KLAS, SI and MIDAS pose the best choice. High sensitivity is provided by FIRMA while high specificity is achieved by ARH, SplicingCompass and MIDAS.

Chapter 4 encompasses our newly developed network-based approach for the detection of regulatory candidates causal for differential splicing events between a cancer (lymphoma subtype) and a control group (tonsil). While differential splicing events can be detected with several technologies, the underlying causes for aberrant splicing and thus the pathological mechanism usually remains unclear. Yet, the understanding of the latter is essential for the development of therapeutic strategies.

In our approach, we assessed all differential splicing events between each lymphoma subtype and the control group. These altered entities were then integrated with potential regulators, i.e. splicing factors from a curated human SF database, into an expression correlation network. More precisely, we constructed two networks for each subtype, a cancerous and a control network, each containing the same nodes but varying edges based on the expression correlation in the respective group samples. We then assessed betweenness centrality for the SFs in both networks and ranked the SFs based on their difference in the cancerous and control network. This ranked list of SFs contains candidates with potential regulatory implications in the differential splicing events observed.

To assess the plausibility of our approach we used differentially expressed SFs. As they are demonstrably altered between conditions, we expected them to rank highly amongst our differentially central SFs. To this end, we determine the enrichment of the differentially expressed SFs amongst the differentially central SFs. Three out of six comparisons showed a significant enrichment. Our approach led to subtype as well as lymphoma specific candidates. To elucidate their role and potential involvement in cancer, we extensively searched the literature from which we obtained coherent findings, i.e. the known involvement of several candidates in other cancer types. Furthermore, we encountered new candidates, such as TRA2A, for which no disease-related publications could be found. As the origin of perturbation leading to aberrant splicing is not on the expression level for differentially central candidates, we investigated different potential areas of impact. First, we mined TRANSFAC, a database providing associations between transcription factors and the targets they control. Here, we found a regulatory relationship between the microRNA miR-133b and PTBP2, a SF differentially central in our approach. Expression analysis of this microRNA could confirm a down-regulation in all lymphoma subtypes investigated. Second, we assessed genomic alterations based on the relative amount of SNPs in lymphoma samples compared to other cancer cell lines from a published data set. Several SF showed an increased amount of SNPs in lymphoma compared to other cancer cell lines.

Chapter 5 covers the multi-level framework we developed for the comparison of RNA sequencing and exon array data with respect to the prediction of differential splicing events. Since RNA sequencing is gradually replacing microarrays, an assessment of comparability of both technologies is indispensable. While several studies exists for the comparison on gene level, work on the differential splicing level is rare and lacks explanatory power.

To ensure high comprehensibility and comparability, we approached the task two-fold. First, we implemented multiple comparison levels, i.e., not only the level of differential splicing, but also several preceding levels to elucidate the origin of potential incomparability. Second, we aimed at improving differential splicing comparability by the use of two DS detection methods applicable to both, RNA sequencing and exon array data to avoid method inherent bias [Raghavachari et al., 2012].

Our multi-level approach comprises comparisons on the probe level, the exon level and the gene level as well as on the level of differential splicing. According to the specific level, we assess expression correlation, expression distribution, fold change correlation and overlap in differentially expressed entities. To ensure comparability of entities on every level, we filter those entities who could be unambiguously mapped. On the probe level, for instance, probe sequences are mapped to the genome, to detect uniquely mapping probes and assess the RNA sequencing based expression specifically for the region interrogated by the respective probe. Two methods, ARH and SplicingCompass, are selected for the detection of differential splicing from RNA sequencing and exon array data.

Rather low expression correlation is observed throughout the levels, partly due to diverging scales in low expression between technologies and a high variation especially on the probe level. Aggregation of entities on higher levels improves result concordance with respect to fold change correlation. The comparison of differential splicing events results in several significant overlaps on data not corrected for multiple testing for the DLBCL data set. While the technology (RNA sequencing vs. exon array) does have an effect on the comparability of predicted differential splicing events for the DLBCL data set, the impact of the DS detection method used is also important according to our evaluation. For the GBM data set, which differs from the DLBCL data set in terms of technology (paired end vs. single end) and different RNA extraction time points, not only the uncorrected, but also the corrected overlaps were significant.

The main achievements of our work can be thus summarized as follows.

- By comparing DS methods for exon arrays according to their performance on different data settings, we could establish a link between influencial global data factors and the performance of the methods based on accuracy, sensitivity and specificity. This insight provides researchers with the prerequisites to choose methods based on their data basis and their study aim.

- We developed a network-based method for the prediction of regulatory elements causal for splicing changes observed between two conditions such as disease and control. The main advantage of our method is the prediction of causal regulatory candidates carrying potential alterations on several levels, i.e. also on other than the transcription level, yet, by only using transcriptomic data.

- We presented a newly developed, comprehensive multi-level framework for the assessment of comparability of DS events detected from RNA sequencing and exon array data. The application on two different data sets showed several significant overlaps with respect to genes predicted as differentially spliced as well as common

trends in comparability. The use of further data sets might help to elucidate the influence of additional factors such as the use of different sequencing technologies.

6.1 Future Directions

This work sheds light on several aspects of the detection of differential splicing based on different methods and technologies as well as on the uncovering of regulatory elements potentially involved in the aberrant splicing observed. Naturally, not all aspects of interest can be covered in a limited amount of time. Additionally, growing insight often leads to further questions arising in the course of research. This section gives an overview on several aspects deemed worthy to investigate the overall goal of understanding disease-related mechanisms associated with differential splicing to eventually identify potential therapeutic targets.

Machine learning and differential splicing. An interesting approach to investigate the distinctive potential of DS events is classification. Feature selection and classification of cancer subtypes based on differentially spliced exons relative to the control group could, first, reveal DS events which are central in their tumor-promoting role. Second, these splicing patterns might be meaningful for diagnosis.

Besides distinguishing lymphoma subtypes from the tonsil control group, the DS events detectable between lymphoma subtypes have a high potential in elucidating the differing patho-mechanisms. DLBCL, for instance, is known to consist of (at least) three different subtypes, ABC, GCB and others. While gene signatures differentiating between these groups are known, a distinction based on DS patterns would provide an additional, potentially more fine-grained, view on the underlying molecular mechanisms separating the three classes.

Finally, clustering of cancer types based on their splicing pattern might reveal common subclasses over cancer types representing, for instance, susceptibility to the same treatment.

Extending the comparison of methods for DS detection. Our approach on the comparison of DS detection methods using artificial and experimental data substantially elucidates the performance of the methods compared, yet, both data types have drawbacks. The artificial data naturally represents only a subset of the biologically observed splicing events with respect to exon number as well as the amount of differentially spliced exons per gene. A data setting covering further scenarios would highly increase the number of tested data set, and could lead to relevant insights on the susceptibility of the methods to the naturally occurring settings. Although experimental data is of high value for evaluations, the number of data points is rather low and unbalanced. Additional evaluation of methods with confirmed splicing events from for instance tissue databases could provide a more robust assessment of performance. Finally, an indispensable step is the experimental validation of the predicted DS events.

Enrichment of the splicing regulatory network approach. Our SF regulatory approach leverages transcriptomic data by the emergence of knowledge due to a network-based approach. While this enables the prediction of candidates with potential

malfunctions on levels other than the transcriptomic, the approach would profit from an enrichment of information based on other levels alterations can occur on.

This could be either done, by integrating players from other levels such as epigenetics or components of the phosphorylatory machinery directly into the network based on their expression values. Potential candidates for this matter are histone modifying and methylation-associated genes. For influences due to phosphorylation changes kinases and phosphatases can be included. A more sophisticated approach constitutes the integration of analysis results gained on other levels. Two important areas in this context are epigenetic changes and genomic alterations. Modified histones or methylated DNA can explain changes in expression, either for splicing factors or for concrete differential splicing events. Similarly, mutations and other genomic alterations can influence the function and splicing behavior of genes and the resulting proteins as well the cis-regulatory elements controlling the splicing events.

Network correlation based approaches rely on a relatively high number of samples. While expression data based on exon arrays provided necessary group sizes, RNA sequencing based sample sizes in our approach do not offer the possibility of robust correlation network analysis. Yet, RNA sequencing data is highly beneficial in determining accurate splicing events, because of the unbiased nature of the technology. Furthermore, spliced reads provide the ultimate prove for the existence of a splicing event, while exon arrays rely on gene-relative exon expression changes. On that account, an evaluation of the SF network approach based on RNA sequencing data is of high interest and could reveal an even more accurate picture of potentially deregulated actors.

Several network centrality measures exist. In the past few years, these 'basic' measures such as betweenness, degree and closeness centrality have been modified and refined for the detection of differentially important components especially in biological networks. A recent comparison of such DiNA (differential network analysis) methods accentuates the superiority of local and hybrid DiNA methods compared to global approaches [Lichtblau et al., 2016]. Consequently, an adaption of the SF network approach using measures highly ranked in the evaluation of Lichtblau et al [Lichtblau et al., 2016] might improve candidate prediction.

Splicing factors and their targets. It is of high interest to identify regulatory candidates potentially involved in aberrant DS. Yet, to understand the effect and the underlying patho-mechanism of such regulators, the concrete targets of SF have to be identified, i.e. the 'deregulated' SF and the exons they control have to be associated. While several attempts to solve the issue exist [Aittokallio and Schwikowski, 2006, Dai et al., 2011, Chen and Zheng, 2009], this task is non-trivial due to two reasons. First, exons and SF usually share n:m relations, i.e. a SF may control several exons and one exon might be controlled by several SFs. Second, most of the SF play rather a meta-role in the regulation of a certain exon by participating, for instance, in regulatory complexes. Thus, no direct information on binding to specific DNA regions such as cis-regulatory elements can be used.

Investigate new candidates. Several of the identified candidates amongst the SFs are known in the context of cancer. For some, even mechanistic explanations can be found. While these SF can be interpreted as a substantiation of our approach, the

truly interesting candidates are the ones not previously associated with cancer such as TRA2A. Thus, a specific evaluation of these candidates based on different impacting levels such as epigenetics or genomic alterations might shed light on their supposedly unknown role in DS.

Exon based comparison. One of our central aims in the comparison of exon array data to RNA sequencing data related to DS was the warranty of maximal comparability. Hence, we chose two methods applicable to both data types. These methods provide predictions of DS on a gene basis, which naturally impedes the comparison of differentially spliced exons. Nevertheless, such a comparison is of high interest and could be achieved by using different methods for the two technologies.

Appendix

Exon.ID	Gene.ID	Symbol	p-value
3394697	3394660	TRIM29	7.514500e-16
3832868	3832865	NCCRP1	9.272199e-16
3757075	3757050	KRT13	1.059239e-14
3394661	3394660	TRIM29	1.157338e-14
3455883	3455865	KRT4	1.723497e-14
3494641	3494629	SCEL	1.723497e-14
3494642	3494629	SCEL	1.934432e-14
3638433	3638411	RHCG	2.714679e-14
3497228	3497195	CLDN10	3.114437e-14
3455882	3455865	KRT4	9.538484e-14
3455885	3455865	KRT4	9.835962e-14
3113183	3113180	MAL2	1.289319e-13
3455236	3455186	KRT5	2.309053e-13
3497233	3497195	CLDN10	4.920022e-13
3455237	3455186	KRT5	5.209625e-13

Table 6.1: Differential Splicing in ALCL

Exon.ID	Gene.ID	Symbol	p-value
3497228	3497195	CLDN10	5.292302e-22
3832868	3832865	NCCRP1	5.292302e-22
3497233	3497195	CLDN10	1.330475e-21
3394697	3394660	TRIM29	2.571146e-20
3497227	3497195	CLDN10	2.571146e-20
3638433	3638411	RHCG	2.571146e-20
3455875	3455865	KRT4	2.160791e-19
3394661	3394660	TRIM29	3.188063e-19
3757075	3757050	KRT13	3.954294e-19
3455885	3455865	KRT4	4.675862e-19
3455883	3455865	KRT4	5.722255e-19
3757098	3757078	KRT15	5.722255e-19
2893795	2893794	DSP	6.888564e-19
3494642	3494629	SCEL	1.077927e-18
3494641	3494629	SCEL	1.187819e-18

Table 6.2: Differential Splicing in DLBCL

Exon.ID	Gene.ID	Symbol	p-value
3497228	3497195	CLDN10	7.410631e-18
3455883	3455865	KRT4	1.696985e-17
3757075	3757050	KRT13	1.154848e-16
3497233	3497195	CLDN10	1.410789e-16
3494641	3494629	SCEL	1.560160e-16
3832868	3832865	NCCRP1	1.580196e-16
3394661	3394660	TRIM29	2.675233e-16
3783501	3783481	DSG3	2.675233e-16
3832869	3832865	NCCRP1	2.675233e-16
3757098	3757078	KRT15	3.092309e-16
3497227	3497195	CLDN10	3.412170e-16
3394697	3394660	TRIM29	3.589900e-16
3493550	3493543	KLF5	3.589900e-16
3638433	3638411	RHCG	3.589900e-16
3394662	3394660	TRIM29	8.679099e-16

Table 6.3: Differential Splicing in CLL

Exon.ID	Gene.ID	Symbol	p-value
3394697	3394660	TRIM29	2.069926e-20
3638433	3638411	RHCG	5.676567e-20
3497228	3497195	CLDN10	8.795846e-20
3394661	3394660	TRIM29	5.743939e-19
3757075	3757050	KRT13	6.185971e-19
3832868	3832865	NCCRP1	1.186134e-18
3494641	3494629	SCEL	2.799307e-18
3497227	3497195	CLDN10	3.774077e-18
3494642	3494629	SCEL	8.137291e-18
3497233	3497195	CLDN10	8.210589e-18
3455883	3455865	KRT4	9.811480e-18
3455885	3455865	KRT4	9.811480e-18
3455875	3455865	KRT4	1.160549e-17
3455236	3455186	KRT5	1.813686e-17
3455882	3455865	KRT4	3.623111e-17

Table 6.4: Differential Splicing in FL

Exon.ID	Gene.ID	Symbol	p-value
3757098	3757078	KRT15	8.986443e-14
3497233	3497195	CLDN10	6.175881e-13
3497227	3497195	CLDN10	2.069670e-12
3497228	3497195	CLDN10	1.182240e-11
3832869	3832865	NCCRP1	1.182240e-11
3394661	3394660	TRIM29	2.890094e-11
3455885	3455865	KRT4	2.890094e-11
3455883	3455865	KRT4	4.596178e-11
3757075	3757050	KRT13	5.530383e-11
3394697	3394660	TRIM29	6.881004e-11
3394698	3394660	TRIM29	2.235965e-10
3497220	3497195	CLDN10	3.184868e-10
3832868	3832865	NCCRP1	3.275280e-10
3873104	3873102	ZCCHC3	3.839708e-10
3493553	3493543	KLF5	4.630996e-10

Table 6.5: Differential Splicing in MCL

Exon.ID	Gene.ID	Symbol	p-value
3497228	3497195	CLDN10	1.345192e-12
3832868	3832865	NCCRP1	1.345192e-12
3497233	3497195	CLDN10	1.942635e-12
3638433	3638411	RHCG	2.076993e-12
3494642	3494629	SCEL	2.202101e-12
3455875	3455865	KRT4	8.534811e-12
3394697	3394660	TRIM29	1.231007e-11
3455883	3455865	KRT4	1.392877e-11
3394698	3394660	TRIM29	1.925563e-11
3455882	3455865	KRT4	1.925563e-11
3494641	3494629	SCEL	2.435742e-11
3638455	3638411	RHCG	3.525309e-11
3497227	3497195	CLDN10	4.614446e-11
3455885	3455865	KRT4	5.517594e-11
3394661	3394660	TRIM29	1.712838e-10

Table 6.6: Differential Splicing in PTCL

GO-BP-Term	PValue
GO:0007155~cell adhesion	9.651084e-20
GO:0022610~biological adhesion	1.026714e-19
GO:0007398~ectoderm development	2.454536e-17

GO:0008544~epidermis development	3.566239e-17
GO:0030855~epithelial cell differentiation	1.173162e-11
GO:0060429~epithelium development	1.241629e-09
GO:0030198~extracellular matrix organization	1.752353e-09
GO:0043062~extracellular structure organization	1.284571e-08
GO:0009913~epidermal cell differentiation	2.617371e-08
GO:0031589~cell-substrate adhesion	7.707046e-08
GO:0042060~wound healing	1.206190e-07
GO:0030216~keratinocyte differentiation	1.240062e-07
GO:0016337~cell-cell adhesion	1.762735e-07
GO:0009611~response to wounding	7.693936e-07
GO:0007160~cell-matrix adhesion	2.141154e-06
GO:0018149~peptide cross-linking	3.451386e-06
GO:0043588~skin development	6.830011e-06
GO:0044259~multicellular organismal macromolecule metabolic process	1.029626e-05
GO:0001568~blood vessel development	1.493244e-05
GO:0001944~vasculature development	1.990398e-05
GO:0044236~multicellular organismal metabolic process	2.991119e-05
GO:0031424~keratinization	7.217721e-05
GO:0032963~collagen metabolic process	8.755150e-05
GO:0030199~collagen fibril organization	1.043685e-04
GO:0001501~skeletal system development	2.904855e-04
GO:0006928~cell motion	3.714555e-04
GO:0022404~molting cycle process	6.272225e-04
GO:0022405~hair cycle process	6.272225e-04
GO:0001942~hair follicle development	6.272225e-04
GO:0042303~molting cycle	7.002548e-04
GO:0042633~hair cycle	7.002548e-04
GO:0048730~epidermis morphogenesis	7.161899e-04
GO:0048514~blood vessel morphogenesis	8.168157e-04
GO:0010033~response to organic substance	9.379530e-04
GO:0060537~muscle tissue development	1.147810e-03
GO:0048545~response to steroid hormone stimulus	1.410180e-03
GO:0016477~cell migration	2.278116e-03
GO:0032964~collagen biosynthetic process	2.664497e-03
GO:0007517~muscle organ development	2.820849e-03
GO:0001525~angiogenesis	3.323506e-03
GO:0009725~response to hormone stimulus	3.365417e-03
GO:0031069~hair follicle morphogenesis	3.613188e-03
GO:0035295~tube development	3.828357e-03
GO:0048870~cell motility	5.347931e-03
GO:0051674~localization of cell	5.347931e-03
GO:0007044~cell-substrate junction assembly	6.288796e-03

GO:0018108~peptidyl-tyrosine phosphorylation	7.039514e-03
GO:0048732~gland development	7.369119e-03
GO:0009719~response to endogenous stimulus	7.902488e-03
GO:0018212~peptidyl-tyrosine modification	8.180662e-03
GO:0044243~multicellular organismal catabolic process	8.901341e-03
GO:0048745~smooth muscle tissue development	9.179835e-03

Table 6.7: David result for ALCL

GO-BP-Term	PValue
GO:0008544~epidermis development	1.244998e-18
GO:0007398~ectoderm development	1.771867e-18
GO:0007155~cell adhesion	3.044954e-11
GO:0022610~biological adhesion	3.143533e-11
GO:0009913~epidermal cell differentiation	4.079993e-11
GO:0030855~epithelial cell differentiation	4.682975e-10
GO:0030216~keratinocyte differentiation	1.354111e-08
GO:0060429~epithelium development	3.232999e-08
GO:0016337~cell-cell adhesion	1.103497e-06
GO:0007160~cell-matrix adhesion	2.091860e-05
GO:0042060~wound healing	4.545833e-05
GO:0018149~peptide cross-linking	4.693114e-05
GO:0031589~cell-substrate adhesion	5.197877e-05
GO:0009611~response to wounding	6.668294e-05
GO:0001942~hair follicle development	1.004836e-04
GO:0022404~molting cycle process	1.004836e-04
GO:0022405~hair cycle process	1.004836e-04
GO:0042633~hair cycle	1.173426e-04
GO:0042303~molting cycle	1.173426e-04
GO:0048730~epidermis morphogenesis	4.223051e-04
GO:0007156~homophilic cell adhesion	6.812048e-04
GO:0031424~keratinization	8.526912e-04
GO:0043588~skin development	8.658579e-04
GO:0048732~gland development	8.754943e-04
GO:0007584~response to nutrient	1.180796e-03
GO:0031069~hair follicle morphogenesis	1.336101e-03
GO:0033273~response to vitamin	1.691239e-03
GO:0006928~cell motion	1.745394e-03
GO:0043586~tongue development	2.547936e-03
GO:0042552~myelination	3.016158e-03
GO:0021675~nerve development	3.850792e-03
GO:0051045~negative regulation of membrane protein ectodomain proteolysis	4.000575e-03

GO:0007272~ensheathment of neurons	4.692407e-03
GO:0008366~axon ensheathment	4.692407e-03
GO:0033189~response to vitamin A	4.692407e-03
GO:0008637~apoptotic mitochondrial changes	8.464369e-03
GO:0035270~endocrine system development	9.484398e-03

Table 6.8: David result for DLBCL

GO-BP-Term	PValue
GO:0008544~epidermis development	2.617887e-14
GO:0007398~ectoderm development	4.233548e-14
GO:0030855~epithelial cell differentiation	8.389885e-09
GO:0007155~cell adhesion	4.823272e-08
GO:0022610~biological adhesion	4.971303e-08
GO:0060429~epithelium development	7.335057e-08
GO:0009913~epidermal cell differentiation	1.271995e-07
GO:0030216~keratinocyte differentiation	3.064921e-07
GO:0016337~cell-cell adhesion	1.013295e-06
GO:0001942~hair follicle development	7.735445e-05
GO:0022404~molting cycle process	7.735445e-05
GO:0022405~hair cycle process	7.735445e-05
GO:0042633~hair cycle	9.221465e-05
GO:0042303~molting cycle	9.221465e-05
GO:0006928~cell motion	9.975445e-05
GO:0031069~hair follicle morphogenesis	3.661198e-04
GO:0043586~tongue development	3.751442e-04
GO:0031175~neuron projection development	5.504453e-04
GO:0048666~neuron development	6.729412e-04
GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	7.415843e-04
GO:0048732~gland development	7.626238e-04
GO:0007160~cell-matrix adhesion	9.356956e-04
GO:0030030~cell projection organization	9.491384e-04
GO:0048812~neuron projection morphogenesis	1.106097e-03
GO:0030182~neuron differentiation	1.224145e-03
GO:0048730~epidermis morphogenesis	1.411054e-03
GO:0018149~peptide cross-linking	1.698309e-03
GO:0033273~response to vitamin	1.832977e-03
GO:0007156~homophilic cell adhesion	1.850382e-03
GO:0000904~cell morphogenesis involved in differentiation	1.951537e-03
GO:0031589~cell-substrate adhesion	1.960116e-03
GO:0042060~wound healing	2.492372e-03

GO:0007229~integrin-mediated signaling pathway	2.674283e-03
GO:0007584~response to nutrient	3.224240e-03
GO:0031424~keratinization	3.308120e-03
GO:0051402~neuron apoptosis	3.541409e-03
GO:0007167~enzyme linked receptor protein signaling pathway	3.659886e-03
GO:0043484~regulation of RNA splicing	4.308142e-03
GO:0016049~cell growth	4.759593e-03
GO:0048858~cell projection morphogenesis	4.808829e-03
GO:0032989~cellular component morphogenesis	5.345851e-03
GO:0048667~cell morphogenesis involved in neuron differentiation	5.776437e-03
GO:0000902~cell morphogenesis	5.855899e-03
GO:0007409~axonogenesis	6.879216e-03
GO:0016477~cell migration	7.119917e-03
GO:0042981~regulation of apoptosis	7.326437e-03
GO:0032990~cell part morphogenesis	7.381775e-03
GO:0007507~heart development	7.451537e-03
GO:0007242~intracellular signaling cascade	8.159232e-03
GO:0042127~regulation of cell proliferation	8.353677e-03
GO:0043067~regulation of programmed cell death	8.600802e-03
GO:0010941~regulation of cell death	9.189182e-03

Table 6.9: David result for CLL

GO-BP-Term	PValue
GO:0007398~ectoderm development	3.664522e-20
GO:0008544~epidermis development	5.445687e-19
GO:0030855~epithelial cell differentiation	1.292044e-15
GO:0060429~epithelium development	1.347844e-14
GO:0007155~cell adhesion	6.914246e-13
GO:0022610~biological adhesion	7.249979e-13
GO:0009913~epidermal cell differentiation	1.502323e-11
GO:0030216~keratinocyte differentiation	1.008115e-09
GO:0016337~cell-cell adhesion	1.951863e-09
GO:0042060~wound healing	5.066139e-06
GO:0007156~homophilic cell adhesion	1.490032e-05
GO:0048732~gland development	1.979916e-05
GO:0048730~epidermis morphogenesis	5.587490e-05
GO:0018149~peptide cross-linking	6.822180e-05
GO:0031424~keratinization	8.348452e-05
GO:0031069~hair follicle morphogenesis	2.624236e-04
GO:0009611~response to wounding	2.949778e-04

GO:0001942~hair follicle development	7.065423e-04
GO:0022405~hair cycle process	7.065423e-04
GO:0022404~molting cycle process	7.065423e-04
GO:0043586~tongue development	7.327775e-04
GO:0042303~molting cycle	7.885236e-04
GO:0042633~hair cycle	7.885236e-04
GO:0007160~cell-matrix adhesion	8.121753e-04
GO:0033273~response to vitamin	8.978528e-04
GO:0048729~tissue morphogenesis	1.058597e-03
GO:0007229~integrin-mediated signaling pathway	1.224475e-03
GO:0031589~cell-substrate adhesion	1.433116e-03
GO:0007584~response to nutrient	2.773635e-03
GO:0042552~myelination	3.878236e-03
GO:0008366~axon ensheathment	5.577025e-03
GO:0007272~ensheathment of neurons	5.577025e-03
GO:0048878~chemical homeostasis	6.896339e-03
GO:0050678~regulation of epithelial cell proliferation	7.256797e-03
GO:0042592~homeostatic process	9.509398e-03
GO:0042445~hormone metabolic process	9.610745e-03

Table 6.10: David result for FL

GO-BP-Term	PValue
GO:0007398~ectoderm development	5.613947e-17
GO:0008544~epidermis development	1.174290e-16
GO:0030855~epithelial cell differentiation	1.292047e-13
GO:0060429~epithelium development	1.064189e-11
GO:0007155~cell adhesion	1.086191e-09
GO:0022610~biological adhesion	1.124393e-09
GO:0009913~epidermal cell differentiation	1.673166e-09
GO:0030216~keratinocyte differentiation	1.025312e-08
GO:0016337~cell-cell adhesion	1.525707e-08
GO:0042060~wound healing	3.431587e-05
GO:0022404~molting cycle process	1.872121e-04
GO:0022405~hair cycle process	1.872121e-04
GO:0001942~hair follicle development	1.872121e-04
GO:0031424~keratinization	2.096723e-04
GO:0042303~molting cycle	2.096723e-04
GO:0042633~hair cycle	2.096723e-04
GO:0007156~homophilic cell adhesion	2.769314e-04
GO:0043586~tongue development	3.144916e-04
GO:0048732~gland development	3.396212e-04
GO:0031069~hair follicle morphogenesis	1.713021e-03

GO:0001501~skeletal system development	2.683653e-03
GO:0048730~epidermis morphogenesis	3.843757e-03
GO:0018149~peptide cross-linking	4.304688e-03
GO:0042692~muscle cell differentiation	4.591194e-03
GO:0007160~cell-matrix adhesion	5.644139e-03
GO:0048878~chemical homeostasis	6.085985e-03
GO:0050801~ion homeostasis	6.358789e-03
GO:0019725~cellular homeostasis	6.907428e-03
GO:0042592~homeostatic process	7.606665e-03
GO:0009611~response to wounding	8.147573e-03
GO:0031589~cell-substrate adhesion	8.427509e-03
GO:0048729~tissue morphogenesis	8.429813e-03
GO:0007569~cell aging	8.462436e-03

Table 6.11: David result for MCL

GO-BP-Term	PValue
GO:0007398~ectoderm development	2.923868e-14
GO:0008544~epidermis development	1.136894e-13
GO:0030855~epithelial cell differentiation	7.609231e-11
GO:0060429~epithelium development	4.155874e-09
GO:0009913~epidermal cell differentiation	1.848859e-06
GO:0030216~keratinocyte differentiation	1.540328e-05
GO:0007155~cell adhesion	1.639677e-04
GO:0022610~biological adhesion	1.666692e-04
GO:0022405~hair cycle process	3.651095e-04
GO:0022404~molting cycle process	3.651095e-04
GO:0001942~hair follicle development	3.651095e-04
GO:0042303~molting cycle	4.000526e-04
GO:0042633~hair cycle	4.000526e-04
GO:0031069~hair follicle morphogenesis	4.610159e-04
GO:0042060~wound healing	9.370350e-04
GO:0048730~epidermis morphogenesis	1.056117e-03
GO:0043588~skin development	1.638697e-03
GO:0016337~cell-cell adhesion	1.845824e-03
GO:0035295~tube development	2.111749e-03
GO:0031424~keratinization	5.096930e-03
GO:0007160~cell-matrix adhesion	5.949153e-03
GO:0031589~cell-substrate adhesion	8.326389e-03
GO:0030324~lung development	8.623702e-03
GO:0030323~respiratory tube development	9.557112e-03

Table 6.12: David result for PTCL

Transcript Exon Array ID	Symbol
2325526	SRRM1
2328465	KHDRBS1
2335671	ELAVL4
2348060	PTBP2
2418000	ZRANB2
2464499	HNRNPU
2487639	PCBP1
2548871	HNRPLL
2558511	TIA1
2593670	SF3B1
2648141	MBNL1
2709062	TRA2B
2775463	HNRNPD
2775562	HNRPDL
2877141	HNRNPA0
2890148	HNRNPH1
2935475	QKI
2959039	KHDRBS2
2963407	SYNCRIP
3041519	TRA2A
3041550	TRA2A
3042421	HNRNPA2B1
3107548	ESRP1
3117384	KHDRBS3
3201784	ELAVL2
3205033	YBX1
3249738	HNRNPH3
3286286	HNRNPF
3309629	TIAL1
3377044	SF1
3416036	PCBP2
3416483	HNRNPA1
3543411	RBM25
3558745	NOVA1
3656904	FUS
3815834	DAZAP1
3819543	HNRNPM
3848689	ELAVL1
3850960	ELAVL3
3861617	HNRNPL
3865807	NOVA2
3959203	RBFOX2

3994100	FMR1
2320048	TARDBP
2396415	TARDBP
2406064	SFPQ
2421782	RBMX
2454661	HNRNPH1
2588827	HNRNPA3
2594627	HNRNPA1
2622469	RBM5
2893130	HNRNPA1
3212294	HNRNPK
3336422	RBM4
3378411	RBM4
3465593	HNRNPA1
3696226	ESRP2
3847814	KHSRP
3944273	RBFOX2
3969802	HNRPDL
3984779	HNRNPH2
2902804	C2
2904248	SNRPC

Table 6.13: Splicing Factors derived from SpliceAid-F and mapped to our data.

Sample IDs
TCGA-32-2638-01A-01R-1850-01
TCGA-12-1597-01B-01R-1849-01
TCGA-19-2624-01A-01R-1850-01
TCGA-41-2571-01A-01R-1850-01
TCGA-06-0675-11A-32R-A36H-07*
TCGA-06-0678-11A-32R-A36H-07*
TCGA-06-0680-11A-32R-A36H-07*
TCGA-06-0681-11A-41R-A36H-07*

Table 6.14: Samples for glioblastoma multiforme and four organ-specific control samples(*) derived from TCGA [Tomczak et al., 2015].

Bibliography

[Affymetrix, 2005a] Affymetrix (2005a). Alternative Transcript Analysis Methods for Exon Arrays. http://www.affymetrix.com/support/technical/whitepapers/exon_alt_transcript_analysis_whitepaper.pdf.

[Affymetrix, 2005b] Affymetrix (2005b). Exon Probeset Annotations and Transcript Cluster Groupings. http://www.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf.

[Affymetrix, 2005c] Affymetrix (2005c). Gene signal estimates from exon arrays. http://www.affymetrix.com/support/technical/whitepapers/exon_gene_signal_estimate_whitepaper.pdf.

[Affymetrix, 2005d] Affymetrix (2005d). GeneChip Exon Array Design. http://www.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf.

[Affymetrix, 2006] Affymetrix (2006). Human Exon 1.0 ST Array and WT Sense Target Labeling Assay for Genome-Wide, Exon-Level Expression Analysis. http://www.affymetrix.com/support/technical/technotes/human_exon_wt_target_technote.pdf.

[Agafonov et al., 2011] Agafonov, D. E., Deckert, J., Wolf, E., Odenwälder, P., Bessonov, S., Will, C. L., Urlaub, H., and Lührmann, R. (2011). Semi-quantitative proteomic analysis of the human spliceosome via a novel two-dimensional gel electrophoresis method. *Molecular and cellular biology*, pages MCB–05266.

[Agatheeswaran et al., 2012] Agatheeswaran, S., Singh, S., Biswas, S., Biswas, G., Pattnayak, N. C., and Chakraborty, S. (2012). Bcr-abl mediated repression of mir-223 results in the activation of mef2c and ptbp2 in chronic myeloid leukemia. *Leukemia*, 27(7):1578–1580.

[Aittokallio and Schwikowski, 2006] Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics*, 7(3):243–255.

[Akgul et al., 2004] Akgul, C., Moulding, D., and Edwards, S. (2004). Alternative splicing of bcl-2-related genes: functional consequences and potential therapeutic applications. *Cellular and Molecular Life Sciences CMLS*, 61(17):2189–2199.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

[American society for microbiology, 2017] American society for microbiology, C. M. R. (2017). Photolitography . `http://cmr.asm.org/content/22/4/611/F3.expansion.html`.

[Anders et al., 2014] Anders, S., Pyl, P. T., and Huber, W. (2014). Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638.

[Anders et al., 2012] Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from rna-seq data. *Genome research*, 22(10):2008–2017.

[Andrews et al., 2010] Andrews, S. et al. (2010). Fastqc: A quality control tool for high throughput sequence data. *Reference Source*.

[Aschoff et al., 2013] Aschoff, M., Hotz-Wagenblatt, A., Glatting, K.-H., Fischer, M., Eils, R., and König, R. (2013). Splicingcompass: differential splicing detection using rna-seq data. *Bioinformatics*, page btt101.

[Bäckhed, 2012] Bäckhed, F. (2012). Host responses to the human microbiome. *Nutrition reviews*, 70(suppl 1):S14–S17.

[Barash et al., 2010] Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., and Frey, B. J. (2010). Deciphering the splicing code. *Nature*, 465(7294):53–59.

[Barberan-Soler and Zahler, 2008] Barberan-Soler, S. and Zahler, A. M. (2008). Alternative splicing regulation during c. elegans development: splicing factors as regulated targets. *PLoS genetics*, 4(2):e1000001.

[Baylin and Jones, 2011] Baylin, S. B. and Jones, P. A. (2011). A decade of exploring the cancer epigenome—biological and translational implications. *Nature Reviews Cancer*, 11(10):726–734.

[Beier and Hoheisel, 2000] Beier, M. and Hoheisel, J. D. (2000). Production by quantitative photolithographic synthesis of individually quality checked dna microarrays. *Nucleic Acids Research*, 28(4):e11–e11.

[Bellare et al., 2006] Bellare, P., Kutach, A. K., Rines, A. K., Guthrie, C., and Sontheimer, E. J. (2006). Ubiquitin binding by a variant jab1/mpn domain in the essential pre-mrna splicing factor prp8p. *Rna*, 12(2):292–302.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

[Bernstein et al., 2007] Bernstein, B. E., Meissner, A., and Lander, E. S. (2007). The mammalian epigenome. *Cell*, 128(4):669–681.

[Birikh et al., 2003] Birikh, K. R., Sklan, E. H., Shoham, S., and Soreq, H. (2003). Interaction of "readthrough" acetylcholinesterase with rack1 and pkcβii correlates with intensified fear-induced conflict behavior. *Proceedings of the National Academy of Sciences*, 100(1):283–288.

[Birzele et al., 2008] Birzele, F., Csaba, G., and Zimmer, R. (2008). Alternative splicing and protein structure evolution. *Nucleic acids research*, 36(2):550–558.

[Bisognin et al., 2014] Bisognin, A., Pizzini, S., Perilli, L., Esposito, G., Mocellin, S., Nitti, D., Zanovello, P., Bortoluzzi, S., and Mandruzzato, S. (2014). An integrative framework identifies alternative splicing events in colorectal cancer development. *Molecular oncology*, 8(1):129–141.

[Black, 2000] Black, D. L. (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, 103(3):367–370.

[Black, 2003] Black, D. L. (2003). Mechanisms of alternative pre-messenger rna splicing. *Annual review of biochemistry*, 72(1):291–336.

[Bolger et al., 2014] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, page btu170.

[Bosch et al., 2015] Bosch, L. J., de Wit, M., Hiemstra, A. C., Piersma, S., Pham, T., Oudgenoeg, G., Scheffer, G., Mongera, S., Komor, M., Droste, J. T. S., et al. (2015). Stool proteomics reveals novel candidate biomarkers for colorectal cancer screening.

[Bottomly et al., 2011] Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., and Hitzemann, R. (2011). Evaluating gene expression in c57bl/6j and dba/2j mouse striatum using rna-seq and microarrays. *PloS one*, 6(3):e17820.

[Bradford et al., 2010] Bradford, J. R., Hey, Y., Yates, T., Li, Y., Pepper, S. D., and Miller, C. J. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC genomics*, 11(1):282.

[Brown et al., 2012] Brown, S. J., Stoilov, P., and Xing, Y. (2012). Chromatin and epigenetic regulation of pre-mrna processing. *Human molecular genetics*, page dds353.

[Buck and Lieb, 2004] Buck, M. J. and Lieb, J. D. (2004). Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360.

[Buckanovich et al., 1996] Buckanovich, R. J., Yang, Y., and Darnell, R. B. (1996). The onconeural antigen nova-1 is a neuron-specific rna-binding protein, the activity of which is inhibited by paraneoplastic antibodies. *The Journal of neuroscience*, 16(3):1114–1122.

[Cáceres and Kornblihtt, 2002] Cáceres, J. F. and Kornblihtt, A. R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. *TRENDS in Genetics*, 18(4):186–193.

[Calarco et al., 2011] Calarco, J. A., Zhen, M., and Blencowe, B. J. (2011). Networking in a global world: establishing functional connections between neural splicing regulators and their target transcripts. *Rna*, 17(5):775–791.

[Calon et al., 2015] Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Momblona, X., Iglesias, M., Sevillano, M., Palomo-Ponce, S., Tauriello, D. V., Byrom, D., et al. (2015). Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nature genetics*, 47(4):320–329.

[Campo et al., 1999] Campo, E., Raffeld, M., and Jaffe, E. S. (1999). Mantle-cell lymphoma. In *Seminars in hematology*, volume 36, pages 115–127. [Sheboygan, Wis.]: Grune & Stratton,[c1964-.

[Cariaso and Lennon, 2012] Cariaso, M. and Lennon, G. (2012). Snpedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic acids research*, 40(D1):D1308–D1312.

[Chang et al., 2007] Chang, Y.-F., Imam, J. S., and Wilkinson, M. F. (2007). The nonsense-mediated decay rna surveillance pathway. *Annu. Rev. Biochem.*, 76:51–74.

[Chen and Zheng, 2009] Chen, L. and Zheng, S. (2009). Studying alternative splicing regulatory networks through partial correlation analysis. *Genome Biol*, 10(1):R3.

[Chen et al., 2014] Chen, S., Zhang, J., Duan, L., Zhang, Y., Li, C., Liu, D., Ouyang, C., Lu, F., and Liu, X. (2014). Identification of hnrnp m as a novel biomarker for colorectal carcinoma by quantitative proteomics. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 306(5):G394–G403.

[Choudhury et al., 2014] Choudhury, R., Roy, S. G., Tsai, Y. S., Tripathy, A., Graves, L. M., and Wang, Z. (2014). The splicing activator dazap1 integrates splicing control into mek/erk-regulated cell proliferation and migration. *Nature communications*, 5.

[Clark et al., 2002] Clark, T. A., Sugnet, C. W., and Ares, M. (2002). Genomewide analysis of mrna processing in yeast using splicing-specific microarrays. *Science*, 296(5569):907–910.

[Cline et al., 2005] Cline, M. S., Blume, J., Cawley, S., Clark, T. A., Hu, J.-S., Lu, G., Salomonis, N., Wang, H., and Williams, A. (2005). Anosva: a statistical method for detecting splice variation from expression data. *Bioinformatics*, 21(suppl 1):i107–i115.

[Cohen et al., 2008] Cohen, A. A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M., Issaeva, I., Sigal, A., Milo, R., Cohen-Saidon, C., Liron, Y., Kam, Z., et al. (2008). Dynamic proteomics of individual cancer cells in response to a drug. *science*, 322(5907):1511–1516.

[Collins, 1999] Collins, F. S. (1999). Medical and societal consequences of the human genome project. *New England Journal of Medicine*, 341(1):28–37.

[Collins et al., 2003] Collins, F. S., Morgan, M., and Patrinos, A. (2003). The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290.

[Conway et al., 2003] Conway, T. et al. (2003). Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. *Molecular microbiology*, 47(4):879–889.

[Dai et al., 2011] Dai, C., Li, W., Liu, J., and Zhou, X. J. (2011). Systematic reconstruction of splicing regulatory modules by integrating many rna-seq datasets. In *Systems Biology (ISB), 2011 IEEE International Conference on*, pages 267–273. IEEE.

[D'Alessandro et al., 2008] D'Alessandro, V., Muscarella, L. A., Copetti, M., Zelante, L., Carella, M., and Vendemiale, G. (2008). Molecular detection of neuron-specific elav-like-positive cells in the peripheral blood of patients with small-cell lung cancer. *Analytical Cellular Pathology*, 30(4):291–297.

[Dapas et al., 2016] Dapas, M., Kandpal, M., Bi, Y., and Davuluri, R. V. (2016). Comparative evaluation of isoform-level gene expression estimation algorithms for rna-seq and exon-array platforms. *Briefings in bioinformatics*, page bbw016.

[Dardenne et al., 2012] Dardenne, E., Pierredon, S., Driouch, K., Gratadou, L., Lacroix-Triki, M., Espinoza, M. P., Zonta, E., Germann, S., Mortada, H., Villemin, J.-P., et al. (2012). Splicing switch of an epigenetic regulator by rna helicases promotes tumor-cell invasiveness. *Nature structural & molecular biology*, 19(11):1139–1146.

[De La Grange et al., 2010] De La Grange, P., Gratadou, L., Delord, M., Dutertre, M., and Auboeuf, D. (2010). Splicing factor and exon profiling across human tissues. *Nucleic acids research*, 38(9):2825–2838.

[Deininger et al., 2005] Deininger, M., Buchdunger, E., and Druker, B. J. (2005). The development of imatinib as a therapeutic agent for chronic myeloid leukemia. *Blood*, 105(7):2640–2653.

[Del Fabbro et al., 2013] Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on illumina ngs data analysis. *PLoS One*, 8(12):e85024.

[Dittmar et al., 2012] Dittmar, K. A., Jiang, P., Park, J. W., Amirikian, K., Wan, J., Shen, S., Xing, Y., and Carstens, R. P. (2012). Genome-wide determination of a broad

esrp-regulated posttranscriptional network by high-throughput sequencing. *Molecular and cellular biology*, 32(8):1468–1482.

[Djebali et al., 2012] Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

[Draghici et al., 2006] Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. (2006). Reliability and reproducibility issues in dna microarray measurements. *TRENDS in Genetics*, 22(2):101–109.

[Dredge et al., 2005] Dredge, B. K., Stefani, G., Engelhard, C. C., and Darnell, R. B. (2005). Nova autoregulation reveals dual functions in neuronal splicing. *The EMBO journal*, 24(8):1608–1620.

[Durinck et al., 2009] Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184–1191.

[Dvinge et al., 2016] Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R. K. (2016). Rna splicing factors as oncoproteins and tumour suppressors. *Nature Reviews Cancer*, 16(7):413–430.

[Edgar et al., 2002] Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210.

[Erickson et al., 2012] Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of crohn's disease. *PloS one*, 7(11):e49138.

[Estrada, 2006] Estrada, E. (2006). Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1):35–40.

[Fackenthal and Godley, 2008] Fackenthal, J. D. and Godley, L. A. (2008). Aberrant rna splicing and its functional consequences in cancer cells. *Disease models & mechanisms*, 1(1):37–42.

[Faustino and Cooper, 2003] Faustino, N. A. and Cooper, T. A. (2003). Pre-mrna splicing and human disease. *Genes & development*, 17(4):419–437.

[Fay and Shaw, 2010] Fay, M. P. and Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: the interval r package. *Journal of Statistical Software*, 36(2).

[Feinberg et al., 2006] Feinberg, A. P., Ohlsson, R., and Henikoff, S. (2006). The epigenetic progenitor origin of human cancer. *Nature reviews genetics*, 7(1):21–33.

[Feng et al., 2013] Feng, Y., Niu, L., Wei, W., Zhang, W., Li, X., Cao, J., and Zhao, S. (2013). A feedback circuit between mir-133 and the erk1/2 pathway involving an exquisite mechanism for regulating myoblast proliferation and differentiation. *Cell death & disease*, 4(11):e934.

[Fisher, 1922] Fisher, R. A. (1922). On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, pages 87–94.

[Fortier et al., 2013] Fortier, A.-M., Asselin, E., and Cadrin, M. (2013). Keratin 8 and 18 loss in epithelial cancer cells increases collective cell migration and cisplatin sensitivity through claudin1 up-regulation. *Journal of Biological Chemistry*, 288(16):11555–11571.

[Frankish et al., 2015] Frankish, A., Uszczynska, B., Ritchie, G. R., Gonzalez, J. M., Pervouchine, D., Petryszak, R., Mudge, J. M., Fonseca, N., Brazma, A., Guigo, R., et al. (2015). Comparison of gencode and refseq gene annotation and the impact of reference geneset on variant effect prediction. *BMC genomics*, 16(8):S2.

[Freeman, 1977] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

[French et al., 2007] French, P. J., Peeters, J., Horsman, S., Duijm, E., Siccama, I., Van Den Bent, M. J., Luider, T. M., Kros, J. M., van der Spek, P., and Smitt, P. A. S. (2007). Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. *Cancer Research*, 67(12):5635–5642.

[Fu et al., 2013] Fu, R.-H., Liu, S.-P., Huang, S.-J., Chen, H.-J., Chen, P.-R., Lin, Y.-H., Ho, Y.-C., Chang, W.-L., Tsai, C.-H., Shyu, W.-C., et al. (2013). Aberrant alternative splicing events in parkinson's disease. *Cell transplantation*, 22(4):653–661.

[Fu and Ares Jr, 2014] Fu, X.-D. and Ares Jr, M. (2014). Context-dependent control of alternative splicing by rna-binding proteins. *Nature Reviews Genetics*, 15(10):689–701.

[Gardina et al., 2006] Gardina, P. J., Clark, T. A., Shimada, B., Staples, M. K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., et al. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC genomics*, 7(1):1.

[Geiger et al., 2012] Geiger, T., Madden, S. F., Gallagher, W. M., Cox, J., and Mann, M. (2012). Proteomic portrait of human breast cancer progression identifies novel prognostic markers. *Cancer research*, 72(9):2428–2439.

[Ghigna et al., 2005] Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P. M., Green, M. R., Riva, S., and Biamonti, G. (2005). Cell motility is controlled by sf2/asf through alternative splicing of the ron protooncogene. *Molecular cell*, 20(6):881–890.

[Ghigna et al., 2008] Ghigna, C., Valacca, C., and Biamonti, G. (2008). Alternative splicing and tumor progression. *Current genomics*, 9(8):556.

[Giudice et al., 2014] Giudice, J., Xia, Z., Wang, E. T., Scavuzzo, M. A., Ward, A. J., Kalsotra, A., Wang, W., Wehrens, X. H., Burge, C. B., Li, W., et al. (2014). Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nature communications*, 5.

[Giulietti et al., 2012] Giulietti, M., Piva, F., D'Antonio, M., De Meo, P. D., Paoletti, D., Castrignanò, T., D'Erchia, A. M., Picardi, E., Zambelli, F., Principato, G., et al. (2012). Spliceaid-f: a database of human splicing factors and their rna-binding sites. *Nucleic acids research*, page gks997.

[Glinsky et al., 2005] Glinsky, G. V., Berezovska, O., and Glinskii, A. B. (2005). Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *Journal of Clinical Investigation*, 115(6):1503.

[Golde et al., 1990] Golde, T. E., Estus, S., Usiak, M., Younkin, L. H., and Younkin, S. G. (1990). Expression of β amyloid protein precursor mrnas: recognition of a novel alternatively spliced form and quantitation in alzheimer's disease using pcr. *Neuron*, 4(2):253–267.

[Golub et al., 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537.

[Goodwin et al., 2015] Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M., and McCombie, W. R. (2015). Oxford nanopore sequencing and de novo assembly of a eukaryotic genome. *BioRxiv*, page 013490.

[Graveley, 2001] Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *TRENDS in Genetics*, 17(2):100–107.

[Greenblum et al., 2012] Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, 109(2):594–599.

[Griffith et al., 2010] Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C., Pugh, T. J., et al. (2010). Alternative expression analysis by rna sequencing. *Nature methods*, 7(10):843–847.

[Grosso et al., 2008] Grosso, A. R., Gomes, A. Q., Barbosa-Morais, N. L., Caldeira, S., Thorne, N. P., Grech, G., Von Lindern, M., and Carmo-Fonseca, M. (2008). Tissue-specific splicing factor gene expression signatures. *Nucleic acids research*, 36(15):4823–4832.

[Gunderson et al., 1997] Gunderson, S. I., Vagner, S., Polycarpou-Schwarz, M., and Mattaj, I. W. (1997). Involvement of the carboxyl terminus of vertebrate poly (a) polymerase in u1a autoregulation and in the coupling of splicing and polyadenylation. *Genes & development*, 11(6):761–773.

[Hall et al., 2011] Hall, J., Leong, H. S., Armenoult, L., Newton, G., Valentine, H. R., Irlam, J. J., Möller-Levet, C., Sikand, K. A., Pepper, S. D., Miller, C. J., et al. (2011). Exon-array profiling unlocks clinically and biologically relevant gene signatures from formalin-fixed paraffin-embedded tumour samples. *British journal of cancer*, 104(6):971–981.

[Hamosh et al., 2005] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514–D517.

[Handelsman, 2004] Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, 68(4):669–685.

[Haque and Oberdoerffer, 2014] Haque, N. and Oberdoerffer, S. (2014). Chromatin and splicing. *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*, pages 97–113.

[Harrow et al., 2012] Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774.

[Hartwell et al., 2006] Hartwell, L., Mankoff, D., Paulovich, A., Ramsey, S., and Swisher, E. (2006). Cancer biomarkers: a systems approach. *Nature biotechnology*, 24(8):905–908.

[Hayden, 2014] Hayden, E. C. (2014). The $1,000 genome. *Nature*, 507(7492):294–295.

[He et al., 2014] He, Q., He, Q., Liu, X., Wei, Y., Shen, S., Hu, X., Li, Q., Peng, X., Wang, L., and Yu, L. (2014). Genome-wide prediction of cancer driver genes based on snp and cancer snv data. *American journal of cancer research*, 4(4):394.

[Heinzen et al., 2008] Heinzen, E. L., Ge, D., Cronin, K. D., Maia, J. M., Shianna, K. V., Gabriel, W. N., Welsh-Bohmer, K. A., Hulette, C. M., Denny, T. N., and Goldstein, D. B. (2008). Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS biology*, 6(12):e1000001.

[Hope and Murray, 2011] Hope, N. R. and Murray, G. I. (2011). The expression profile of rna-binding proteins in primary and metastatic colorectal cancer: relationship of heterogeneous nuclear ribonucleoproteins with prognosis. *Human pathology*, 42(3):393–402.

[Horiguchi et al., 2011] Horiguchi, K., Sakamoto, K., Koinuma, D., Semba, K., Inoue, A., Inoue, S., Fujii, H., Yamaguchi, A., Miyazawa, K., Miyazono, K., et al. (2011). Tgf-β drives epithelial-mesenchymal transition through δef1-mediated downregulation of esrp. *Oncogene*, 31(26):3190–3201.

[Hu et al., 2010] Hu, G., Chen, D., Li, X., Yang, K., Wang, H., and Wu, W. (2010). mir-133b regulates the met proto-oncogene and inhibits the growth of colorectal cancer cells in vitro and in vivo. *Cancer biology & therapy*, 10(2):190–197.

[Hu et al., 2001] Hu, G. K., Madore, S. J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., and Wang, Y. (2001). Predicting splice variant from dna chip expression data. *Genome Research*, 11(7):1237–1245.

[Hu et al., 2008] Hu, S., Arellano, M., Boontheung, P., Wang, J., Zhou, H., Jiang, J., Elashoff, D., Wei, R., Loo, J. A., and Wong, D. T. (2008). Salivary proteomics for oral cancer biomarker discovery. *Clinical Cancer Research*, 14(19):6246–6252.

[Hu et al., 2015] Hu, Z., Scott, H. S., Qin, G., Zheng, G., Chu, X., Xie, L., Adelson, D. L., Oftedal, B. E., Venugopal, P., Babic, M., et al. (2015). Revealing missing human protein isoforms based on ab initio prediction, rna-seq and proteomics. *Scientific reports*, 5.

[Huang et al., 2008] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57.

[Iannone and Valcárcel, 2013] Iannone, C. and Valcárcel, J. (2013). Chromatin's thread to alternative splicing regulation. *Chromosoma*, 122(6):465–474.

[Illumina, 2017] Illumina (2017). Illumina GenomeAnalyzer IIx. `https://www.illumina.com/content/dam/illumina-marketing/documents/products/brochures/brochure_genome_analyzer.pdf`.

[Indraccolo et al., 2002] Indraccolo, S., Minuzzo, S., Zamarchi, R., Calderazzo, F., Piovan, E., and Amadori, A. (2002). Alternatively spliced forms of igα and igβ prevent b cell receptor expression on the cell surface. *European journal of immunology*, 32(6):1530–1540.

[Irizarry et al., 2003a] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of affymetrix genechip probe level data. *Nucleic acids research*, 31(4):e15–e15.

[Irizarry et al., 2003b] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

[Iwasaki et al., 2009] Iwasaki, R., Kiuchi, H., Ihara, M., Mori, T., Kawakami, M., and Ueda, H. (2009). Trans-splicing as a novel method to rapidly produce antibody fusion proteins. *Biochemical and biophysical research communications*, 384(3):316–321.

[Jensen and Whitehead, 2001] Jensen, L. E. and Whitehead, A. S. (2001). Irak1b, a novel alternative splice variant of interleukin-1 receptor-associated kinase (irak), mediates interleukin-1 signaling and has prolonged stability. *Journal of Biological Chemistry*, 276(31):29037–29044.

[Jentsch, 2011] Jentsch, M. (2011). Alternative Splicing Detection Algorithms for Affymetrix Exon Array Data - Comprehensive Evaluation and New Methods Based on Kullback-Leibler Divergence and Non-Parametric Statistics. Master's thesis, Freie Univesitaet Berlin, Berlin, Germany.

[Jones et al., 2016] Jones, P. A., Issa, J.-P. J., and Baylin, S. (2016). Targeting the cancer epigenome for therapy. *Nature Reviews Genetics*, 17(10):630–641.

[Kar et al., 2005] Kar, A., Kuo, D., He, R., Zhou, J., and Wu, J. Y. (2005). Tau alternative splicing and frontotemporal dementia. *Alzheimer disease and associated disorders*, 19(Suppl 1):S29.

[Karni et al., 2007] Karni, R., de Stanchina, E., Lowe, S. W., Sinha, R., Mu, D., and Krainer, A. R. (2007). The gene encoding the splicing factor sf2/asf is a proto-oncogene. *Nature structural & molecular biology*, 14(3):185–193.

[Kau et al., 2012] Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2012). Human nutrition, the gut microbiome and the immune system. *Nature*, 474:327–336.

[Kazarian and Laird-Offringa, 2011] Kazarian, M. and Laird-Offringa, I. A. (2011). Small-cell lung cancer-associated autoantibodies: potential applications to cancer diagnosis, early detection, and therapy. *Mol Cancer*, 10(1):33.

[Keren et al., 2010] Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355.

[Khaldoyanidi et al., 1996] Khaldoyanidi, S., Achtnich, M., Hehlmann, R., and Zöller, M. (1996). Expression of cd44 variant isoforms in peripheral blood leukocytes in malignant lymphoma and leukemia: inverse correlation between expression and tumor progression. *Leukemia research*, 20(10):839–851.

[Khan et al., 2014] Khan, D. H., Gonzalez, C., Cooper, C., Sun, J.-M., Chen, H. Y., Healy, S., Xu, W., Smith, K. T., Workman, J. L., Leygue, E., et al. (2014). Rna-dependent dynamic histone acetylation regulates mcl1 alternative splicing. *Nucleic acids research*, 42(3):1656–1670.

[Klijn et al., 2014] Klijn, C., Durinck, S., Stawiski, E. W., Haverty, P. M., Jiang, Z., Liu, H., Degenhardt, J., Mayba, O., Gnad, F., Liu, J., et al. (2014). A comprehensive transcriptional portrait of human cancer cell lines. *Nature biotechnology*.

[Koch et al., 2001] Koch, T., Schulz, S., Pfeiffer, M., Klutzny, M., Schröder, H., Kahl, E., and Höllt, V. (2001). C-terminal splice variants of the mouse μ-opioid receptor differ in morphine-induced internalization and receptor resensitization. *Journal of Biological Chemistry*, 276(33):31408–31414.

[Kononen et al., 1998] Kononen, J., Bubendorf, L., Kallionimeni, A., Bärlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., and Kallionimeni, O.-P. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature medicine*, 4(7):844–847.

[Kornblihtt et al., 2013] Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature reviews Molecular cell biology*, 14(3):153–165.

[Korneta et al., 2012] Korneta, I., Magnus, M., and Bujnicki, J. M. (2012). Structural bioinformatics of the human spliceosomal proteome. *Nucleic acids research*, 40(15):7046–7065.

[Koschützki and Schreiber, 2008] Koschützki, D. and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology*, 2:193.

[Krawczak et al., 2007] Krawczak, M., Thomas, N. S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., and Cooper, D. N. (2007). Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mrna splicing. *Human mutation*, 28(2):150.

[Laajala et al., 2009] Laajala, E., Aittokallio, T., Lahesmaa, R., and Elo, L. L. (2009). Probe-level estimation improves the detection of differential splicing in affymetrix exon array studies. *Genome biology*, 10(7):1–12.

[Langer et al., 2010] Langer, W., Sohler, F., Leder, G., Beckmann, G., Seidel, H., Gröne, J., Hummel, M., and Sommer, A. (2010). Exon array analysis using re-defined probe sets results in reliable identification of alternatively spliced genes in non-small cell lung cancer. *BMC genomics*, 11(1):676.

[Le Naour et al., 2001] Le Naour, F., Misek, D. E., Krause, M. C., Deneux, L., Giordano, T. J., Scholl, S., and Hanash, S. M. (2001). Proteomics-based identification of

rs/dj-1 as a novel circulating tumor antigen in breast cancer. *Clinical Cancer Research*, 7(11):3328–3335.

[LeFave et al., 2011] LeFave, C. V., Squatrito, M., Vorlova, S., Rocco, G. L., Brennan, C. W., Holland, E. C., Pan, Y.-X., and Cartegni, L. (2011). Splicing factor hnrnph drives an oncogenic splicing switch in gliomas. *The EMBO journal*, 30(19):4084–4097.

[Lekva et al., 2012] Lekva, T., Berg, J. P., Fougner, S. L., Olstad, O. K., Ueland, T., and Bollerslev, J. (2012). Gene expression profiling identifies esrp1 as a potential regulator of epithelial mesenchymal transition in somatotroph adenomas from a large cohort of patients with acromegaly. *The Journal of Clinical Endocrinology & Metabolism*, 97(8):E1506–E1514.

[Lekva et al., 2013] Lekva, T., Berg, J. P., Lyle, R., Heck, A., Ringstad, G., Olstad, O. K., Michelsen, A. E., Casar-Borota, O., Bollerslev, J., and Ueland, T. (2013). Epithelial splicing regulator protein 1 and alternative splicing in somatotroph adenomas. *Endocrinology*, 154(9):3331–3343.

[Lemma et al., 2013] Lemma, S., Karihtala, P., Haapasaari, K.-M., Jantunen, E., Soini, Y., Bloigu, R., Pasanen, A.-K., Turpeenniemi-Hujanen, T., and Kuittinen, O. (2013). Biological roles and prognostic values of the epithelial–mesenchymal transition-mediating transcription factors twist, zeb1 and slug in diffuse large b-cell lymphoma. *Histopathology*, 62(2):326–333.

[Lenzken et al., 2013] Lenzken, S. C., Loffreda, A., and Barabino, S. M. (2013). Rna splicing: a new player in the dna damage response. *International journal of cell biology*, 2013.

[Li and Wong, 2001] Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li and Homer, 2010] Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483.

[Li et al., 2014] Li, H.-D., Menon, R., Omenn, G. S., and Guan, Y. (2014). The emerging era of genomic data integration for analyzing splice isoform function. *Trends in Genetics*, 30(8):340–347.

[Li and Koromilas, 2001] Li, S. and Koromilas, A. E. (2001). Dominant negative function by an alternatively spliced form of the interferon-inducible protein kinase pkr. *Journal of Biological Chemistry*, 276(17):13881–13890.

[Li et al., 2015] Li, X.-W., Shi, B.-Y., Yang, Q.-L., Wu, J., Wu, H.-M., Wang, Y.-F., Wu, Z.-J., Fan, Y.-M., and Wang, Y.-P. (2015). Epigenetic regulation of cdh1 exon 8 alternative splicing in gastric cancer. *BMC cancer*, 15(1):954.

[Lichtblau et al., 2016] Lichtblau, Y., Zimmermann, K., Haldemann, B., Lenze, D., Hummel, M., and Leser, U. (2016). Comparative assessment of differential network analysis methods. *Briefings in bioinformatics*, page bbw061.

[Lignitto et al., 2014] Lignitto, L., Mattiolo, A., Negri, E., Persano, L., Gianesello, L., Chieco-Bianchi, L., and Calabrò, M. L. (2014). Crosstalk between the mesothelium and lymphomatous cells: insight into the mechanisms involved in the progression of body cavity lymphomas. *Cancer medicine*, 3(1):1–13.

[Liu and Gong, 2008] Liu, F. and Gong, C.-X. (2008). Tau exon 10 alternative splicing and tauopathies. *Mol Neurodegener*, 3(8):1326–1338.

[Liu et al., 2010] Liu, H., Bebu, I., and Li, X. (2010). Microarray probes and probe sets. *Frontiers in bioscience (Elite edition)*, 2:325.

[Liu et al., 2011] Liu, S., Lin, L., Jiang, P., Wang, D., and Xing, Y. (2011). A comparison of rna-seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic acids research*, 39(2):578–588.

[Liu et al., 2013] Liu, Y., Conaway, L., Bethard, J. R., Al-Ayoubi, A. M., Bradley, A. T., Zheng, H., Weed, S. A., and Eblen, S. T. (2013). Phosphorylation of the alternative mrna spf45 by clk1 regulates its splice site utilization, cell migration and invasion. *Nucleic acids research*, page gkt170.

[Lock et al., 2014] Lock, F. E., Rebollo, R., Miceli-Royer, K., Gagnier, L., Kuah, S., Babaian, A., Sistiaga-Poveda, M., Lai, C. B., Nemirovsky, O., Serrano, I., et al. (2014). Distinct isoform of fabp7 revealed by screening for retroelement-activated genes in diffuse large b-cell lymphoma. *Proceedings of the National Academy of Sciences*, 111(34):E3534–E3543.

[Lockstone, 2011] Lockstone, H. E. (2011). Exon array data analysis using affymetrix power tools and r statistical software. *Briefings in bioinformatics*, page bbq086.

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1.

[Luco et al., 2011] Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R., and Misteli, T. (2011). Epigenetics in alternative pre-mrna splicing. *Cell*, 144(1):16–26.

[MacManes, 2014] MacManes, M. D. (2014). On the optimal trimming of high-throughput mrnaseq data. *bioRxiv*, page 000422.

[Malone and Oliver, 2011] Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(1):34.

[Mantione et al., 2014] Mantione, K. J., Kream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M., and Stefano, G. B. (2014). Comparing bioinformatic gene expression profiling methods: microarray and rna-seq. *Medical science monitor basic research*, 20:138–141.

[Mardis, 2008] Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141.

[Marusyk and Polyak, 2010] Marusyk, A. and Polyak, K. (2010). Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1805(1):105–117.

[Marzec et al., 2006] Marzec, M., Kasprzycka, M., Lai, R., Gladden, A. B., Wlodarski, P., Tomczak, E., Nowell, P., DePrimo, S. E., Sadis, S., Eck, S., et al. (2006). Mantle cell lymphoma cells express predominantly cyclin d1a isoform and are highly sensitive to selective inhibition of cdk4 kinase activity. *Blood*, 108(5):1744–1750.

[Matera and Wang, 2014] Matera, A. G. and Wang, Z. (2014). A day in the life of the spliceosome. *Nature reviews Molecular cell biology*, 15(2):108–121.

[Matlin et al., 2005] Matlin, A. J., Clark, F., and Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nature reviews Molecular cell biology*, 6(5):386–398.

[Matter et al., 2002] Matter, N., Herrlich, P., and König, H. (2002). Signal-dependent regulation of splicing via phosphorylation of sam68. *Nature*, 420(6916):691–695.

[Maunakea et al., 2013] Maunakea, A. K., Chepelev, I., Cui, K., and Zhao, K. (2013). Intragenic dna methylation modulates alternative splicing by recruiting mecp2 to promote exon recognition. *Cell research*, 23(11):1256–1269.

[Mazoyer et al., 1998] Mazoyer, S., Puget, N., Perrin-Vidoz, L., Lynch, H. T., Serova-Sinilnikova, O. M., and Lenoir, G. M. (1998). A brca1 nonsense mutation causes exon skipping. *American journal of human genetics*, 62(3):713.

[Merkhofer et al., 2014] Merkhofer, E. C., Hu, P., and Johnson, T. L. (2014). Introduction to cotranscriptional rna splicing. *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*, pages 83–96.

[Meshorer and Soreq, 2006] Meshorer, E. and Soreq, H. (2006). Virtues and woes of ache alternative splicing in stress-related neuropathologies. *Trends in neurosciences*, 29(4):216–224.

[Mischel et al., 2004] Mischel, P. S., Cloughesy, T. F., and Nelson, S. F. (2004). Dna-microarray analysis of brain cancer: molecular classification for therapy. *Nature Reviews Neuroscience*, 5(10):782–792.

[Moulton et al., 2014] Moulton, V. R., Gillooly, A. R., and Tsokos, G. C. (2014). Ubiquitination regulates expression of the serine/arginine-rich splicing factor 1 (srsf1) in normal and systemic lupus erythematosus (sle) t cells. *Journal of Biological Chemistry*, 289(7):4126–4134.

[Nagalakshmi et al., 2008] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349.

[Nakka et al., 2015] Nakka, K. K., Chaudhary, N., Joshi, S., Bhat, J., Singh, K., Chatterjee, S., Malhotra, R., De, A., Santra, M. K., Dilworth, F. J., et al. (2015). Nuclear matrix-associated protein smar1 regulates alternative splicing via hdac6-mediated deacetylation of sam68. *Proceedings of the National Academy of Sciences*, 112(26):E3374–E3383.

[Naro and Sette, 2013] Naro, C. and Sette, C. (2013). Phosphorylation-mediated regulation of alternative splicing in cancer. *International journal of cell biology*, 2013.

[National Cancer Institute, 2017] National Cancer Institute, N. (2017). Lymphatic System . `https://www.cancer.gov/PublishedContent/Images/images/cancer-types/cthp/lymphsystem_male_enlarge.__v2003303802.jpg`.

[Nature, 2010] Nature (2010). Gene Expression. `http://www.nature.com/scitable/topicpage/gene-expression-14121669`.

[Nault et al., 2015] Nault, R., Fader, K. A., and Zacharewski, T. (2015). Rna-seq versus oligonucleotide array assessment of dose-dependent tcdd-elicited hepatic gene expression in mice. *BMC genomics*, 16(1):1.

[Nekrutenko and Taylor, 2012] Nekrutenko, A. and Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672.

[Nilsen and Graveley, 2010] Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463.

[Novello et al., 2013] Novello, C., Pazzaglia, L., Cingolani, C., Conti, A., Quattrini, I., Manara, M. C., Tognon, M., Picci, P., and Benassi, M. S. (2013). mirna expression profile in human osteosarcoma: role of mir-1 and mir-133b in proliferation and cell cycle control. *International journal of oncology*, 42(2):667–675.

[Odibat and Reddy, 2011] Odibat, O. and Reddy, C. K. (2011). Ranking differential genes in co-expression networks. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 350–354. ACM.

[Odibat and Reddy, 2012] Odibat, O. and Reddy, C. K. (2012). Ranking differential hubs in gene co-expression networks. *Journal of bioinformatics and computational biology*, 10(01).

[Oltean and Bates, 2014] Oltean, S. and Bates, D. (2014). Hallmarks of alternative splicing in cancer. *Oncogene*, 33(46):5311–5318.

[Paccione et al., 2008] Paccione, R. J., Miyazaki, H., Patel, V., Waseem, A., Gutkind, J. S., Zehner, Z. E., and Yeudall, W. A. (2008). Keratin down-regulation in vimentin-positive cancer cells is reversible by vimentin rna interference, which inhibits growth and motility. *Molecular cancer therapeutics*, 7(9):2894–2903.

[Pajares et al., 2007] Pajares, M. J., Ezponda, T., Catena, R., Calvo, A., Pio, R., and Montuenga, L. M. (2007). Alternative splicing: an emerging topic in molecular and clinical oncology. *The lancet oncology*, 8(4):349–357.

[Patel and Steitz, 2003] Patel, A. A. and Steitz, J. A. (2003). Splicing double: insights from the second spliceosome. *Nature Reviews Molecular Cell Biology*, 4(12):960–970.

[Piekielko-Witkowska et al., 2010] Piekielko-Witkowska, A., Wiszomirska, H., Wojcicka, A., Poplawski, P., Boguslawska, J., Tanski, Z., and Nauman, A. (2010). Disturbed expression of splicing factors in renal cancer affects alternative splicing of apoptosis regulators, oncogenes, and tumor suppressors. *PLoS One*, 5(10):e13690.

[Pruitt and Maglott, 2001] Pruitt, K. D. and Maglott, D. R. (2001). Refseq and locuslink: Ncbi gene-centered resources. *Nucleic acids research*, 29(1):137–140.

[Purdom et al., 2008] Purdom, E., Simpson, K. M., Robinson, M. D., Conboy, J., Lapuk, A., and Speed, T. P. (2008). Firma: a method for detection of alternative splicing from exon array data. *Bioinformatics*, 24(15):1707–1714.

[Qu et al., 2010] Qu, K., Yesnik, A. M., and Ortoleva, P. J. (2010). Alternative splicing regulatory network reconstruction from exon array data. *Journal of theoretical biology*, 263(4):471–480.

[Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

[Qureshi and Mehler, 2012] Qureshi, I. A. and Mehler, M. F. (2012). Emerging roles of non-coding rnas in brain evolution, development, plasticity and disease. *Nature Reviews Neuroscience*, 13(8):528–541.

[Raghavachari et al., 2012] Raghavachari, N., Barb, J., Yang, Y., Liu, P., Woodhouse, K., Levy, D., O'Donnell, C. J., Munson, P. J., and Kato, G. J. (2012). A systematic comparison and evaluation of high density exon arrays and rna-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC medical genomics*, 5(1):1.

[Rakyan et al., 2011] Rakyan, V. K., Down, T. A., Balding, D. J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541.

[Rasche and Herwig, 2010] Rasche, A. and Herwig, R. (2010). Arh: predicting splice variants from genome-wide data with modified entropy. *Bioinformatics*, 26(1):84–90.

[Rasche et al., 2014] Rasche, A., Lienhard, M., Yaspo, M.-L., Lehrach, H., and Herwig, R. (2014). Arh-seq: identification of differential splicing in rna-seq data. *Nucleic acids research*, 42(14):e110–e110.

[Ratti et al., 2008] Ratti, A., Fallini, C., Colombrita, C., Pascale, A., Laforenza, U., Quattrone, A., and Silani, V. (2008). Post-transcriptional regulation of neuro-oncological ventral antigen 1 by the neuronal rna-binding proteins elav. *Journal of biological chemistry*, 283(12):7531–7541.

[Rauch et al., 2010] Rauch, J., O'Neill, E., Mack, B., Matthias, C., Munz, M., Kolch, W., and Gires, O. (2010). Heterogeneous nuclear ribonucleoprotein h blocks mst2-mediated apoptosis in cancer cells by regulating a-raf transcription. *Cancer research*, 70(4):1679–1688.

[Rehm et al., 2014] Rehm, A., Gätjen, M., Gerlach, K., Scholz, F., Mensen, A., Gloger, M., Heinig, K., Lamprecht, B., Mathas, S., Bégay, V., et al. (2014). Dendritic cell-mediated survival signals in eμ-myc b-cell lymphoma depend on the transcription factor c/ebpβ. *Nature communications*, 5.

[Reinke et al., 2012] Reinke, L. M., Xu, Y., and Cheng, C. (2012). Snail represses the splicing regulator epithelial splicing regulatory protein 1 to promote epithelial-mesenchymal transition. *Journal of Biological Chemistry*, 287(43):36435–36442.

[Relógio et al., 2005] Relógio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R. B., and Valcárcel, J. (2005). Alternative splicing microarrays reveal functional expression of neuron-specific regulators in hodgkin lymphoma cells. *Journal of Biological Chemistry*, 280(6):4779–4784.

[Reuter et al., 2015] Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597.

[Revil et al., 2010] Revil, T., Gaffney, D., Dias, C., Majewski, J., and Jerome-Majewska, L. A. (2010). Alternative splicing is frequent during early embryonic development in mouse. *BMC genomics*, 11(1):399.

[Ritchie et al., 2015] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, page gkv007.

[Rodrigo-Domingo et al., 2013] Rodrigo-Domingo, M., Waagepetersen, R., Bødker, J. S., Falgreen, S., Kjeldsen, M. K., Johnsen, H. E., Dybkær, K., and Bøgsted, M. (2013). Reproducible probe-level analysis of the affymetrix exon 1.0 st array with r/bioconductor. *Briefings in bioinformatics*, page bbt011.

[Rosenwald et al., 2003] Rosenwald, A., Wright, G., Wiestner, A., Chan, W. C., Connors, J. M., Campo, E., Gascoyne, R. D., Grogan, T. M., Muller-Hermelink, H. K., Smeland, E. B., et al. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer cell*, 3(2):185–197.

[Ross et al., 2013] Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol*, 14(5):R51.

[Sakurai et al., 2001] Sakurai, Y., Onishi, Y., Tanimoto, Y., and KIZAKI, H. (2001). Novel protein kinase c δ isoform insensitive to caspase-3. *Biological and Pharmaceutical Bulletin*, 24(9):973–977.

[Salles et al., 1993] Salles, G., Zain, M., Jiang, W., Boussiotis, V. A., and Shipp, M. (1993). Alternatively spliced cd44 transcripts in diffuse large-cell lymphomas: characterization and comparison with normal activated b cells and epithelial malignancies. *Blood*, 82(12):3539–3547.

[Salton et al., 2014] Salton, M., Voss, T. C., and Misteli, T. (2014). Identification by high-throughput imaging of the histone methyltransferase ehmt2 as an epigenetic regulator of vegfa alternative splicing. *Nucleic acids research*, 42(22):13662–13673.

[Sammeth et al., 2008] Sammeth, M., Foissac, S., and Guigó, R. (2008). A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol*, 4(8):e1000147.

[Sanchez-Tillo et al., 2014] Sanchez-Tillo, E., Fanlo, L., Siles, L., Montes-Moreno, S., Moros, A., Chiva-Blanch, G., Estruch, R., Martinez, A., Colomer, D., Győrffy, B., et al. (2014). The emt activator zeb1 promotes tumor growth and determines differential response to chemotherapy in mantle cell lymphoma. *Cell Death & Differentiation*, 21(2):247–257.

[Sanidas et al., 2014] Sanidas, I., Polytarchou, C., Hatziapostolou, M., Ezell, S. A., Kottakis, F., Hu, L., Guo, A., Xie, J., Comb, M. J., Iliopoulos, D., et al. (2014). Phosphoproteomics screen reveals akt isoform-specific signals linking rna processing to lung cancer. *Molecular cell*, 53(4):577–590.

[Schena et al., 1995] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467.

[Schwerk and Schulze-Osthoff, 2005] Schwerk, C. and Schulze-Osthoff, K. (2005). Regulation of apoptosis by alternative pre-mrna splicing. *Molecular cell*, 19(1):1–13.

[Shah and Pallas, 2009] Shah, S. H. and Pallas, J. A. (2009). Identifying differential exon splicing using linear models and correlation coefficients. *BMC bioinformatics*, 10(1):1.

[Shankland et al., 2012] Shankland, K. R., Armitage, J. O., and Hancock, B. W. (2012). Non-hodgkin lymphoma. *The Lancet*, 380(9844):848–857.

[Shapiro et al., 2011] Shapiro, I. M., Cheng, A. W., Flytzanis, N. C., Balsamo, M., Condeelis, J. S., Oktay, M. H., Burge, C. B., and Gertler, F. B. (2011). An emt–driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS genetics*, 7(8):e1002218.

[Shen and Laird, 2013] Shen, H. and Laird, P. W. (2013). Interplay between the cancer genome and epigenome. *Cell*, 153(1):38–55.

[Shin et al., 2011] Shin, K.-H., Kim, R. H., Yu, B., Kang, M. K., Elashoff, D., Christensen, R., Pucar, A., and Park, N.-H. (2011). Expression and mutation analysis of heterogeneous nuclear ribonucleoprotein g in human oral cancer. *Oral oncology*, 47(11):1011–1016.

[Shindo et al., 2013] Shindo, Y., Nozaki, T., Saito, R., and Tomita, M. (2013). Computational analysis of associations between alternative splicing and histone modifications. *FEBS letters*, 587(5):516–521.

[Shipitsin et al., 2014] Shipitsin, M., Small, C., Choudhury, S., Giladi, E., Friedlander, S., Nardone, J., Hussain, S., Hurley, A., Ernst, C., Huang, Y., et al. (2014). Identification of proteomic biomarkers predicting prostate cancer aggressiveness and lethality despite biopsy-sampling error. *British journal of cancer*, 111(6):1201–1212.

[Slotta-Huspenina et al., 2012] Slotta-Huspenina, J., Koch, I., de Leval, L., Keller, G., Klier, M., Bink, K., Kremer, M., Raffeld, M., Fend, F., and Quintanilla-Martinez, L. (2012). The impact of cyclin d1 mrna isoforms, morphology and p53 in mantle cell lymphoma: p53 alterations and blastoid morphology are strong predictors of a high proliferation index. *haematologica*, 97(9):1422–1430.

[Smith and Valcárcel, 2000] Smith, C. W. and Valcárcel, J. (2000). Alternative pre-mrna splicing: the logic of combinatorial control. *Trends in biochemical sciences*, 25(8):381–388.

[Smyth, 2005] Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

[Snyder et al., 2010] Snyder, M., Du, J., and Gerstein, M. (2010). Personal genome sequencing: current approaches and challenges. *Genes & development*, 24(5):423–431.

[Sorek and Cossart, 2010] Sorek, R. and Cossart, P. (2010). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics*, 11(1):9–16.

[Srebrow and Kornblihtt, 2006] Srebrow, A. and Kornblihtt, A. R. (2006). The connection between splicing and cancer. *Journal of cell science*, 119(13):2635–2641.

[Srinivasan et al., 2005] Srinivasan, K., Shiue, L., Hayes, J. D., Centers, R., Fitzwater, S., Loewen, R., Edmondson, L. R., Bryant, J., Smith, M., Rommelfanger, C., et al. (2005). Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, 37(4):345–359.

[Staal et al., 2003] Staal, F., van der Burg, M., Wessels, L., Barendregt, B., Baert, M., van den Burg, C., Van Huffel, C., Langerak, A., van der Velden, V., Reinders, M., et al. (2003). Dna microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-b acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia*, 17(7):1324–1332.

[Stamm, 2008] Stamm, S. (2008). Regulation of alternative splicing by reversible protein phosphorylation. *Journal of Biological Chemistry*, 283(3):1223–1227.

[Stauder et al., 1995] Stauder, R., Eisterer, W., Thaler, J., and Gunthert, U. (1995). Cd44 variant isoforms in non-hodgkin's lymphoma: a new independent prognostic factor. *Blood*, 85(10):2885–2899.

[Stickel et al., 2009] Stickel, J. S., Weinzierl, A. O., Hillen, N., Drews, O., Schuler, M. M., Hennenlotter, J., Wernet, D., Müller, C. A., Stenzl, A., Rammensee, H.-G., et al. (2009). Hla ligand profiles of primary renal cell carcinoma maintained in metastases. *Cancer immunology, immunotherapy*, 58(9):1407–1417.

[Sveen et al., 2015] Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R., and Skotheim, R. (2015). Aberrant rna splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene*.

[Terpe et al., 1994] Terpe, H.-J., Koopmann, R., Imhof, B. A., and Günthert, U. (1994). Expression of integrins and cd44 isoforms in non-hodgkin's lymphomas: Cd44 variant isoforms are preferentially expressed in high-grade malignant lymphomas. *The Journal of pathology*, 174(2):89–100.

[Thomas et al., 2014] Thomas, P., Durek, P., Solt, I., Klinger, B., Witzel, F., Schulthess, P., Mayer, Y., Tikk, D., Blüthgen, N., and Leser, U. (2014). Computer-assisted curation of a human regulatory core network from the biological literature. *Bioinformatics*, page btu795.

[Thorsen et al., 2008] Thorsen, K., Sørensen, K. D., Brems-Eskildsen, A. S., Modin, C., Gaustadnes, M., Hein, A.-M. K., Kruhøffer, M., Laurberg, S., Borre, M., Wang, K., et al. (2008). Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Molecular & Cellular Proteomics*, 7(7):1214–1224.

[Tibshirani et al., 2002] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

[Ting et al., 2006] Ting, A. H., McGarvey, K. M., and Baylin, S. B. (2006). The cancer epigenome—components and functional correlates. *Genes & development*, 20(23):3215–3231.

[Tomczak et al., 2015] Tomczak, K., Czerwinska, P., Wiznerowicz, M., et al. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, 19(1A):A68–A77.

[Trapnell et al., 2012] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578.

[Tzankov et al., 2003] Tzankov, A., Pehrs, A., Zimpfer, A., Ascani, S., Lugli, A., Pileri, S., and Dirnhofer, S. (2003). Prognostic significance of cd44 expression in diffuse large b cell lymphoma of activated and germinal centre b cell-like types: a tissue microarray analysis of 90 cases. *Journal of clinical pathology*, 56(10):747–752.

[Ueda et al., 2013] Ueda, J., Matsuda, Y., Yamahatsu, K., Uchida, E., Naito, Z., Korc, M., and Ishiwata, T. (2013). Epithelial splicing regulatory protein 1 is a favorable prognostic factor in pancreatic cancer that attenuates pancreatic metastases. *Oncogene*.

[Ule et al., 2006] Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J., and Darnell, R. B. (2006). An rna map predicting nova-dependent splicing regulation. *Nature*, 444(7119):580–586.

[Van De Vijver et al., 2002] Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009.

[Van't Veer et al., 2002] Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536.

[Venables et al., 2013] Venables, J. P., Brosseau, J.-P., Gadea, G., Klinck, R., Prinos, P., Beaulieu, J.-F., Lapointe, E., Durand, M., Thibault, P., Tremblay, K., et al. (2013). Rbfox2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Molecular and cellular biology*, 33(2):396–405.

[Vitting-Seerup et al., 2014] Vitting-Seerup, K., Porse, B. T., Sandelin, A., and Waage, J. (2014). splicer: an r package for classification of alternative splicing and prediction of coding potential from rna-seq data. *BMC bioinformatics*, 15(1):81.

[Wahl et al., 2009] Wahl, M. C., Will, C. L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic rnp machine. *Cell*, 136(4):701–718.

[Wallach-Dayan et al., 2001] Wallach-Dayan, S. B., Grabovsky, V., Moll, J., Sleeman, J., Herrlich, P., Alon, R., and Naor, D. (2001). Cd44-dependent lymphoma cell dissemination: a cell surface cd44 variant, rather than standard cd44, supports in vitro lymphoma cell rolling on hyaluronic acid substrate and its in vivo accumulation in the peripheral lymph nodes. *Journal of cell science*, 114(19):3463–3477.

[Wang et al., 2014a] Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., et al. (2014a). A comprehensive study design reveals treatment-and transcript abundance–dependent concordance between rna-seq and microarray data. *Nature biotechnology*, 32(9):926.

[Wang et al., 2014b] Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., et al. (2014b). The concordance between rna-seq and microarray data depends on chemical treatment and transcript abundance. *Nature biotechnology*, 32(9):926–932.

[Wang et al., 2008] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476.

[Wang et al., 2010] Wang, L., Li, P., and Brutnell, T. P. (2010). Exploring plant transcriptomes using ultra high-throughput sequencing. *Briefings in Functional Genomics*, 9(2):118–128.

[Wang et al., 1999] Wang, X., Bruderer, S., Rafi, Z., Xue, J., Milburn, P. J., Krämer, A., and Robinson, P. J. (1999). Phosphorylation of splicing factor sf1 on ser20 by cgmp-dependent protein kinase regulates spliceosome assembly. *The EMBO journal*, 18(16):4549–4559.

[Wang et al., 2009a] Wang, X., Wang, K., Radovich, M., Wang, Y., Wang, G., Feng, W., Sanford, J. R., and Liu, Y. (2009a). Genome-wide prediction of cis-acting rna elements regulating tissue-specific pre-mrna alternative splicing. *BMC genomics*, 10(1):1.

[Wang and Burge, 2008] Wang, Z. and Burge, C. B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna*, 14(5):802–813.

[Wang et al., 2009b] Wang, Z., Gerstein, M., and Snyder, M. (2009b). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.

[Warf and Berglund, 2010] Warf, M. B. and Berglund, J. A. (2010). Role of rna structure in regulating pre-mrna splicing. *Trends in biochemical sciences*, 35(3):169–178.

[Warzecha et al., 2010] Warzecha, C. C., Jiang, P., Amirikian, K., Dittmar, K. A., Lu, H., Shen, S., Guo, W., Xing, Y., and Carstens, R. P. (2010). An esrp-regulated

splicing programme is abrogated during the epithelial–mesenchymal transition. *The EMBO journal*, 29(19):3286–3300.

[Warzecha et al., 2009a] Warzecha, C. C., Sato, T. K., Nabet, B., Hogenesch, J. B., and Carstens, R. P. (2009a). Esrp1 and esrp2 are epithelial cell-type-specific regulators of fgfr2 splicing. *Molecular cell*, 33(5):591–601.

[Warzecha et al., 2009b] Warzecha, C. C., Shen, S., Xing, Y., and Carstens, R. P. (2009b). The epithelial splicing factors esrp1 and esrp2 positively and negatively regulate diverse types of alternative splicing events. *thyroid*, 13:14.

[Watson et al., 2013] Watson, I. R., Takahashi, K., Futreal, P. A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics*, 14(10):703–718.

[Will and Lührmann, 2011] Will, C. L. and Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor perspectives in biology*, 3(7):a003707.

[Wingender et al., 1996] Wingender, E., Dietze, P., Karas, H., and Knüppel, R. (1996). Transfac: a database on transcription factors and their dna binding sites. *Nucleic acids research*, 24(1):238–241.

[Xi et al., 2008] Xi, L., Feber, A., Gupta, V., Wu, M., Bergemann, A. D., Landreneau, R. J., Litle, V. R., Pennathur, A., Luketich, J. D., and Godfrey, T. E. (2008). Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic acids research*, 36(20):6535–6547.

[Xiang and Li, 2014] Xiang, K.-M. and Li, X.-R. (2014). Mir-133b acts as a tumor suppressor and negatively regulates tbpl1 in colorectal cancer cells. *Asian Pac J Cancer Prev*, 15:3767–72.

[Xing et al., 2008] Xing, Y., Stoilov, P., Kapur, K., Han, A., Jiang, H., Shen, S., Black, D. L., and Wong, W. H. (2008). Mads: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *Rna*, 14(8):1470–1479.

[Xu et al., 2011] Xu, W., Seok, J., Mindrinos, M. N., Schweitzer, A. C., Jiang, H., Wilhelmy, J., Clark, T. A., Kapur, K., Xing, Y., Faham, M., et al. (2011). Human transcriptome array for high-throughput clinical studies. *Proceedings of the National Academy of Sciences*, 108(9):3707–3712.

[Xu et al., 2014] Xu, Y., Gao, X. D., Lee, J.-H., Huang, H., Tan, H., Ahn, J., Reinke, L. M., Peter, M. E., Feng, Y., Gius, D., et al. (2014). Cell type-restricted activity of hnrnpm promotes breast cancer metastasis via regulating alternative splicing. *Genes & development*.

[Yae et al., 2012] Yae, T., Tsuchihashi, K., Ishimoto, T., Motohara, T., Yoshikawa, M., Yoshida, G. J., Wada, T., Masuko, T., Mogushi, K., Tanaka, H., et al. (2012). Alternative splicing of cd44 mrna by esrp1 enhances lung colonization of metastatic cancer cell. *Nature communications*, 3:883.

[Yang and Speed, 2002] Yang, Y. H. and Speed, T. (2002). Design issues for cdna microarray experiments. *Nature Reviews Genetics*, 3(8):579–588.

[Yeo et al., 2004] Yeo, G., Holste, D., Kreiman, G., and Burge, C. B. (2004). Variation in alternative splicing across human tissues. *Genome biology*, 5(10):R74.

[Yu et al., 2007] Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, 3(4):e59.

[Zhang et al., 2015] Zhang, X., Joehanes, R., Chen, B. H., Huan, T., Ying, S., Munson, P. J., Johnson, A. D., Levy, D., and O'Donnell, C. J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nature genetics*, 47(4):345–352.

[Zhang et al., 2014] Zhang, Y.-A., Zhu, J.-M., Yin, J., Tang, W.-Q., Guo, Y.-M., Shen, X.-Z., and Liu, T.-T. (2014). High expression of neuro-oncological ventral antigen 1 correlates with poor prognosis in hepatocellular carcinoma. *PloS one*, 9(3):e90955.

[Zhao et al., 2013a] Zhao, H., Li, M., Li, L., Yang, X., Lan, G., and Zhang, Y. (2013a). Mir-133b is down-regulated in human osteosarcoma and inhibits osteosarcoma cells proliferation, migration and invasion, and promotes apoptosis. *PloS one*, 8(12):e83571.

[Zhao et al., 2014a] Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2014a). Crossmap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007.

[Zhao et al., 2017] Zhao, J.-X., Li, X.-W., Shi, B.-Y., Wang, F., Xu, Z.-R., Meng, H.-L., Su, Y.-Y., Wang, J.-M., Xiao, N., He, Q., et al. (2017). Effect of histone modifications on hmlh1 alternative splicing in gastric cancer. *Tumor Biology*, 39(4):1010428317697546.

[Zhao et al., 2013b] Zhao, Q., Caballero, O. L., Davis, I. D., Jonasch, E., Tamboli, P., Yung, W. A., Weinstein, J. N., Strausberg, R. L., Yao, J., et al. (2013b). Tumor-specific isoform switch of the fibroblast growth factor receptor 2 underlies the mesenchymal and malignant phenotypes of clear cell renal cell carcinomas. *Clinical Cancer Research*, 19(9):2460–2472.

[Zhao et al., 2014b] Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014b). Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1):e78644.

[Zhao et al., 2014c] Zhao, Y., Huang, J., Zhang, L., Qu, Y., Li, J., Yu, B., Yan, M., Yu, Y., Liu, B., and Zhu, Z. (2014c). Mir-133b is frequently decreased in gastric cancer and its overexpression reduces the metastatic potential of gastric cancer cells. *BMC cancer*, 14(1):34.

[Zheng et al., 2009] Zheng, H., Hang, X., Zhu, J., Qian, M., Qu, W., Zhang, C., and Deng, M. (2009). Remas: a new regression model to identify alternative splicing events from exon array data. *BMC bioinformatics*, 10(Suppl 1):S18.

[Zhong et al., 2009] Zhong, X.-Y., Ding, J.-H., Adams, J. A., Ghosh, G., and Fu, X.-D. (2009). Regulation of sr protein phosphorylation and alternative splicing by modulating kinetic interactions of srpk1 with molecular chaperones. *Genes & development*, 23(4):482–495.

[Zhou et al., 2014] Zhou, H.-L., Luo, G., Wise, J. A., and Lou, H. (2014). Regulation of alternative splicing by local histone modifications: potential roles for rna-guided mechanisms. *Nucleic acids research*, 42(2):701–713.

[Zhou et al., 2012] Zhou, Y., Lu, Y., and Tian, W. (2012). Epigenetic features are significantly associated with alternative splicing. *BMC genomics*, 13(1):123.

[Zhou et al., 2002] Zhou, Z., Licklider, L. J., Gygi, S. P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419(6903):182–185.

[Zhu et al., 2001] Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. (2001). Global analysis of protein activities using proteome chips. *science*, 293(5537):2101–2105.

[Zimmermann et al., 2015] Zimmermann, K., Jentsch, M., Rasche, A., Hummel, M., and Leser, U. (2015). Algorithms for differential splicing detection using exon arrays: a comparative assessment. *BMC Genomics*, 16(1):1.

[Zimmermann and Leser, 2010] Zimmermann, K. and Leser, U. (2010). Analysis of affymetrix exon arrays. Technical report, Department for Computer Science, Humboldt-Universität zu Berlin. Technical Report 235.

List of Figures

List of Tables

Selbständigkeitserklärung

Hiermit erkläre ich,

- dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.
- dass ich keinen Doktorgrad im Fach Informatik besitze,
- und dass mir die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät der Humboldt Universität zu Berlin veröffentlicht im amtlichen Mitteilungsblatt Nr. 126/2014 am 18.11.2014 bekannt ist.

Berlin, den 26. Juni 2017 Karin Zimmermann